GREF: GRAPH-GUIDED REFERENCE SELECTION FOR GENERALIZABLE INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frequent product changes make repeated training of industrial anomaly detectors costly, motivating methods that generalize to unseen products using normal reference images. When the available normal pool exceeds the reference budget, detection performance depends on which images are selected. Retrieving the most similar images individually can produce redundant references that provide limited complementary evidence. We propose GRef, a graph-guided framework that combines transferable relational representations with reference selection balancing query relevance and redundancy. It supports both selector-guided detector training and plug-in inference without target-side parameter updates. We establish a generalization bound that accounts for discrete reference selection, relating detection error to selection margins and model complexity. Across six industrial datasets, ResAD++ trained with our selected references achieves an average Image AUROC of 87.94%, outperforming all compared learned detectors. Plug-in integration with UniVAD achieves the best overall average of 93.70%. With detector parameters fixed, GRef also achieves the highest average performance among the compared selection strategies for all four evaluated detectors.

1 INTRODUCTION

Industrial anomaly detection identifies defects that deviate from normal product appearance and is an important component of automated visual inspection. In manufacturing, inspection systems must accommodate new products and variants whose materials, textures, and structures can differ substantially from those previously inspected. Conventional methods typically build product-specific models or normal feature representations (Deng & Li, 2022; Roth et al., 2022; Liu et al., 2023). Applying them to a new product therefore requires collecting normal data and rebuilding or adapting the corresponding detector. Deploying conventional detectors on new products often requires renewed data collection, cleaning, and model training. Repeating this process leads to substantial costs and slows down the deployment of inspection systems for new products. These challenges motivate methods that generalize to unseen products without target-side training.

Recent methods address these challenges by training on source products and using normal images from unseen products as inference-time references (Zhu & Pang, 2024; Yao et al., 2024; Qu et al., 2025). These references provide normal examples of the target product without requiring updates to the learned parameters. Existing methods mainly focus on how to compare a query with the supplied references and use their differences to identify anomalies. However, they typically treat the reference set as a given input, paying less attention to which normal images are most useful for the current query. Normal images can differ in appearance and spatial configuration, so they do not provide equally relevant evidence for every query (Dong et al., 2026). In practice, verified normal images collected during new product setup or routine quality checks form a pool of N reference candidates. Since comparing each query against all N images is computationally expensive, the detector typically selects only $K < N$ references for comparison. This makes the choice of these K references even more important, as different subsets may capture different normal patterns and therefore produce different detection results for the same query. Reference selection thus directly affects overall performance on new unseen target datasets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Fig. 1 illustrates the effect of reference selection on unseen product categories. Even in a simple example of selecting four references from eight normal candidates, different selection strategies lead to substantial performance differences for the same fixed reference-conditioned detector. This highlights the importance of reference selection in anomaly detection. A natural approach is to select references most similar to the query, but this does not explicitly account for redundancy within the set. The selected references may resemble one another and provide overlapping information, limiting the range of normal patterns available to the detector. Effective selection should therefore balance query relevance with complementary information across the selected references. However, identifying such a set is challenging because query labels are unavailable at inference, preventing direct evaluation of its detection utility.

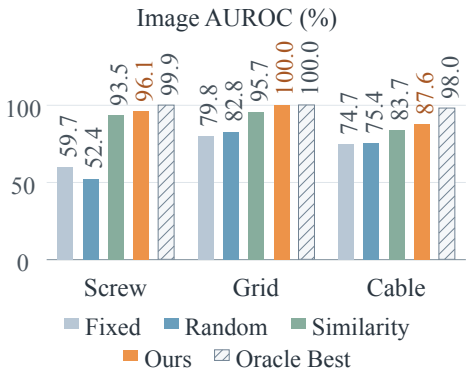

Figure 1: Reference selection (4 of 8). Oracle Best: highest test-set AUROC over all 70 fixed subsets.

Therefore, we propose GRef, a graph-guided framework that integrates query-specific reference selection with anomaly detection. A relational encoder incorporates spatial adjacency and hierarchical containment among image regions to learn transferable representations for reference retrieval. Using these representations, the selector retrieves relevant candidates from the normal pool and selects a compact reference set by balancing query relevance and redundancy. The encoder is pretrained on source normal images and then frozen during selector-guided detector training. The selector also supports plug-in use with an existing detector by replacing its reference input. At inference, references are selected for each query from the target normal pool without parameter updates.

Our contributions are threefold.

- We formulate query-dependent normal-reference selection for generalizable industrial anomaly detection, explicitly distinguishing the available normal pool from the limited reference set used for each query.
- We introduce GRef, which combines a transferable relational encoder with reference selection that balances query relevance and redundancy. The selector supports both detector training and plug-in inference on unseen products without target-side parameter updates.
- We establish a generalization bound for reference-conditioned anomaly detection that accounts for discrete reference selection. The bound relates expected detection error to empirical error, the frequency of small selection margins, and the complexity of the encoder and detector.

Extensive experiments across six industrial anomaly detection datasets demonstrate the generalization ability of GRef. ResAD++ trained with our reference selection achieves an average Image AUROC of 87.94%, outperforming all compared learned detectors. Integrating our selector into UniVAD at inference achieves the best overall average of 93.70%.

2 Related Work

Generalizable Industrial Anomaly Detection Conventional industrial anomaly detection typically builds category-specific models or normal feature representations (Deng & Li, 2022; Roth et al., 2022; Liu et al., 2023), requiring adaptation when new products are introduced. Few-shot and generalizable methods reduce this dependence by using normal references from unseen categories. RegAD (Huang et al., 2022) and FastRecon (Fang et al., 2023) use reference-based registration and feature reconstruction, respectively . InCTRL (Zhu & Pang, 2024) and ResAD(Yao et al., 2024; 2025) learn transferable residual representations , while DictAS formulates anomaly segmentation as dictionary lookup (Qu et al., 2025). HNQ-GAD (Wang et al., 2026) uses residual-guided attention to aggregate query-related normality proxies from support features and localize anomalies in the query. DCP-SFR (Jiang et al., 2026) amplifies early defect cues to guide reconstruction-based localization and refines segmentation boundaries using structural information. UniVAD (Gu et al., 2025) offers a

training-free alternative through component-aware patch matching and graph-enhanced component modeling.

These reference-based methods primarily study anomaly detection given a small normal support set. Although they may adapt feature matching or aggregation to the query, they do not explicitly address which reference images to select from a larger normal pool for each query. Our setting distinguishes the N available normal images from the $K < N$ references used by the detector. This introduces a selection problem that requires both assessing query relevance and accounting for redundancy among references. GRef addresses this problem by constructing a query-dependent reference set, complementing existing reference-conditioned detectors.

Reference Modeling and Selection Beyond detector design, recent studies examine how normal reference evidence is constructed and used. D^{2}4FAD (Dong et al., 2026) learns query-conditioned weights to adjust the contributions of support images in few-shot medical anomaly detection. This highlights that reference value depends on the query, but operates within an existing support set. QueCo (Yun et al., 2026) constructs reliable normal evidence from an unverified test pool through query-conditioned consensus.

Although these methods model relations between queries and reference evidence, selecting a compact reference set also requires considering how references complement one another. Simple strategies include random sampling and nearest-neighbor retrieval with globally pooled WideResNet features (Zagoruyko & Komodakis, 2016). Random sampling ignores query relevance, while global feature similarity may overlook regional structure relevant to normal-reference comparison. CPR (Li et al., 2024) retrieves normal images using block-wise feature histograms before local patch matching. However, both rank candidates individually without explicitly considering redundancy among selected references. This motivates jointly considering reference representations and set composition. GRef learns transferable representations of spatial adjacency and hierarchical containment among image regions to select relevant, less redundant references from a larger normal pool under a fixed reference budget.

3 METHOD

We at first formally define the problem that GRef intends to solve.

Definition 1 (Generalizable Anomaly Detection with Reference Selection). *Given C labeled historical source datasets $\mathcal{D}_s = \{D_j\}_{j=1}^{C}$, where $D_j = \{(\boldsymbol{x}_i^{(j)}, \boldsymbol{y}_i^{(j)})\}_{i=1}^{N_j}$ contains N_j samples, and $N_s = \sum_j N_j$. Each $\boldsymbol{x}_i^{(j)} \in \mathbb{R}^{3\times H\times W}$ is an image and $\boldsymbol{y}_i^{(j)} \in \{0,1\}^{H\times W}$ is its pixel-level anomaly mask, with 1 indicating an anomalous pixel. We consider an unseen target dataset, i.e. not represented in $\mathcal{D}_s$, corresponding to a single product, denoted by $\mathcal{D}_t = (\mathcal{R}_t, \mathcal{Q}_t)$, where $\mathcal{R}_t = \{\boldsymbol{r}_j\}_{j=1}^{N_t}$ is a pool of verified normal reference images available at inference, and $\mathcal{Q}_t = \{\boldsymbol{q}_i\}_{i=1}^{N_q}$ is an unlabeled test set, with $\mathcal{R}_t \cap \mathcal{Q}_t = \varnothing$. Let $\boldsymbol{y}_i^t$ denote the ground-truth anomaly mask of q_i used for evaluation. We also assume the test pair $(\boldsymbol{q}_i, \boldsymbol{y}_i^t)$ is drawn from a distribution P_t.*

Our goal is to learn a reference-selection encoder E_ψ and an anomaly detector f_θ from $\mathcal{D}_s$ to minimize prediction error on unseen target data. For a query $\boldsymbol{q} \in \mathcal{Q}_t$, the encoder guides a selection operator Π_K to construct $\mathcal{S}_q = \Pi_K(\boldsymbol{q}, \mathcal{R}_t; E_\psi)$, where $\mathcal{S}_q \subseteq \mathcal{R}_t$ and $|\mathcal{S}_q| = K < N_t$. This selected subset provides the normal references for f_θ, yielding the prediction $\widehat{\boldsymbol{y}} = f_\theta(\boldsymbol{q}, \Pi_K(\boldsymbol{q}, \mathcal{R}_t; E_\psi))$. The generalization objective is therefore formulated as

$$\min_{\psi,\theta} \quad \mathbb{E}_{(\boldsymbol{q},\boldsymbol{y})\sim P_t}\left[\ell(f_\theta(\boldsymbol{q}, \Pi_K(\boldsymbol{q}, \mathcal{R}_t; E_\psi)), \boldsymbol{y})\right], \tag{1}$$

where ℓ measures the discrepancy between the predicted anomaly map and the ground-truth mask. Both ψ and θ are learned from $\mathcal{D}_s$ and remain fixed during target inference.

In this formulation, reference selection determines which normal patterns are available to the detector for interpreting a query. Effective selection therefore requires identifying relevant references and combining them into an informative set. However, a representation learned on source products may not reliably capture reference relevance on unseen target products. Moreover, with only K references selected, individually relevant images may provide overlapping evidence, leaving other normal patterns needed for comparison unrepresented. These considerations lead to two questions:

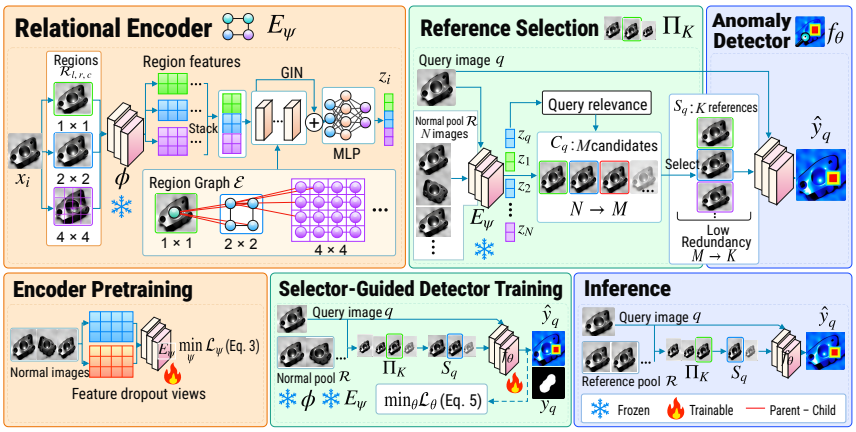

Figure 2: Overview of GRef. The relational encoder E_ψ models spatial adjacency and hierarchical containment among image regions to learn transferable representations for reference retrieval. The selector Π_K retrieves M candidates from N normal images, then selects K references by balancing relevance and redundancy. These references condition f_θ to predict the query's anomaly map. The encoder is pretrained on normal images and frozen during detector training. Inference requires no target-side parameter updates.

Q1 How can we learn an image representation that provides transferable cues for reference relevance on unseen products?

Q2 How can we select K references that are relevant to the query and jointly provide complementary normal evidence?

To address these two questions, we propose transferable relational retrieval and query-adaptive reference set selection in the following, respectively.

3.1 Transferable Relational Retrieval

To learn a transferable representation for reference retrieval, we exploit relations among regions as domain-agnostic structural information. Normal images of the same product often share regional organization despite variations in local appearance. Although specific appearances and layouts vary across products, spatial adjacency and hierarchical relations provide a shared basis for describing this organization across domains. We therefore learn a graph encoder that incorporates this structural information into pretrained regional features.

Relational representation. Modeling relations among regions can help learned representations generalize to new inputs (Battaglia et al., 2018). For reference retrieval on unseen products, we apply this principle to spatial adjacency and containment among image regions. Although product appearance varies, these geometric relations follow the same definitions across products. We therefore incorporate them into regional appearance features to learn a transferable retrieval representation. As shown in Figure 2, we divide an image into a $2^l \times 2^l$ grid at each level $l \in \{0, \dots, J\}$. Let $R_{l,r,c}$ denote the patch region at row r and column c in the grid. When $l = 0$ represents the whole image, and each finer level subdivides every region into four smaller regions. Each region defines a node in the sparse graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with corresponding adjacency matrix $A_\mathcal{E}$, where $\mathcal{V} = \{1, \dots, n\}$ indexes regions by level, row, and column. Bidirectional edges $\mathcal{E}$ connect only horizontal and vertical neighbors within each level and parent–child regions between consecutive levels. Specifically, the parent of $R_{l,r,c}$ for $l > 0$ is $R_{l-1,\lfloor r/2 \rfloor,\lfloor c/2 \rfloor}$. These connections allow each region to exchange information with its spatial neighbors and broader context. To assign node features, we use a frozen feature extractor ϕ to obtain a regional representation $\mathbf{v}_{l,r,c} = \phi(R_{l,r,c}) \in \mathbb{R}^d$. Stacking these features in the fixed node order gives $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n]^\top \in \mathbb{R}^{n \times d}$. For computational efficiency, we approximate independent region encoding by extracting a shared feature map from the full image and applying adaptive average pooling at each grid level. Together, $\mathbf{V}$ and $\mathcal{E}$ provide a structured representation of the image, where node features describe regional appearance and edges encode their spatial relations.

Given this region graph, we use a GIN-based network (Xu et al., 2019) to propagate information along the predefined spatial adjacency and containment relations. With a skip connection, the image representation is

$$z(x) = E_\psi(x) = \mathrm{MLP}\left(\mathrm{vec}\left[\mathbf{V} + \mathrm{GIN}(\mathbf{V}, \mathcal{E})\right]\right) \in \mathbb{R}^d, \tag{2}$$

where ψ denotes the learnable parameters of the GIN and MLP. The operator vec flattens the updated features in a fixed order of level, channel, and spatial position, and the MLP performs nonlinear projection from $\mathbb{R}^{nd}$ to $\mathbb{R}^d$. The resulting representation combines regional appearance with context learned through spatial adjacency and containment.

Relational encoder pretraining. The retrieval and subset selection in Π_K return discrete reference indices, so the detection loss cannot be directly backpropagated through selection to E_ψ. We therefore pretrain E_ψ independently on source normal images without anomaly labels.

Let $\mathcal{D}_s^n = \{\boldsymbol{x}_i\}_{i=1}^{N_s}$ denote all source normal images. For each $\boldsymbol{x}_i$, we independently apply random feature dropout twice to its frozen regional features, producing two perturbed views $\widetilde{\boldsymbol{x}}_i^{(1)}$ and $\widetilde{\boldsymbol{x}}_i^{(2)}$. Their representations are $\widetilde{\boldsymbol{h}}_i^{(1)} = E_\psi(\widetilde{\boldsymbol{x}}_i^{(1)})$, $\widetilde{\boldsymbol{h}}_i^{(2)} = E_\psi(\widetilde{\boldsymbol{x}}_i^{(2)})$. We define the similarity between two perturbed representations as $\widetilde{s}_{ij}^{(1)} = \cos\left(\widetilde{\boldsymbol{h}}_i^{(1)}, \widetilde{\boldsymbol{h}}_j^{(1)}\right), \widetilde{s}_{ij}^{(2)} = \cos\left(\widetilde{\boldsymbol{h}}_i^{(2)}, \widetilde{\boldsymbol{h}}_j^{(2)}\right)$, and $\bar{s}_{ij} = \cos(\mathrm{vec}(\mathbf{V}_i), \mathrm{vec}(\mathbf{V}_j))$, Then, the encoder is pretrained by

$$\min_\psi \mathcal{L}_\psi = \left[\frac{1}{N_s}\sum_{i=1}^{N_s}\left(1 - \cos\left(\widetilde{\boldsymbol{h}}_i^{(1)}, \widetilde{\boldsymbol{h}}_i^{(2)}\right)\right) + \frac{1}{2N_s(N_s-1)}\sum_{k=1}^{2}\sum_{i=1}^{N_s}\sum_{\substack{j=1 \\ j\neq i}}^{N_s}\rho\left(\widetilde{s}_{ij}^{(k)}, \bar{s}_{ij}\right)\right], \tag{3}$$

where ρ is the smooth ℓ_1 loss. In practice, the encoder is optimized within each mini-batch sampled from $\mathcal{D}_s^n$, and the feature-dropout masks are resampled at every optimization step.

The first term encourages consistent representations under independent feature perturbations (Chen et al., 2020). The second requires the similarities between perturbed representations to recover those of the clean frozen regional features (Tung & Mori, 2019). Since these target similarities vary across image pairs, the relational term also prevents a constant representation. Together, the two terms form a graph-conditioned denoising objective. GIN uses spatial neighbors and parent–child regions to compensate for randomly removed feature responses while preserving the cross-image geometry used for retrieval. After pretraining, E_ψ remains fixed during detector training and inference.

3.2 Query-Adaptive Reference Set Selection

The relational representation learned in the previous section allows us to measure reference relevance across different products. However, selecting the K most relevant references independently may produce similar normal examples, leaving the detector with fewer distinct normal patterns. To address the second question, we first retrieve $M > K$ candidates using the learned relational similarity, then select K references from them to maintain query relevance while reducing redundancy.

The candidate set should include references that match the query while leaving enough choices for the final selection. Let $\mathcal{R}$ denote the normal reference pool associated with query $\boldsymbol{q}$, with $N = |\mathcal{R}|$. Let $\mathbf{z}_q = E_\psi(\boldsymbol{q})$ and $\mathbf{z}_i = E_\psi(\boldsymbol{r}_i)$ denote the relational representations of query q and reference r_i, respectively. For each reference $r_i \in \mathcal{R}$, we compute $g_i = \frac{1+\cos(\mathbf{z}_q, \mathbf{z}_i)}{2}$, and an intermediate candidate set $\mathcal{C}_q = \mathrm{Top}\mathrm{M}_{r_i \in \mathcal{R}}\, g_i$, where $K < M < N$ and $\cos(\cdot,\cdot)$ denotes cosine similarity. Keeping $M > K$ relevant candidates preserves query relevance while allowing a diverse set of K references to be selected.

Within this candidate set, relevance scores alone do not account for similar normal patterns shared by different references. We therefore compare each candidate with both the query and the references already selected. Using maximum marginal relevance (MMR) (Carbonell & Goldstein, 1998), we start from $\mathcal{S} = \varnothing$ and repeatedly add

$$r^* = \arg\max_{r_i \in \mathcal{C}_q \setminus \mathcal{S}}\left[\lambda g_i - (1-\lambda)\max_{r_j \in \mathcal{S}}\frac{1+\cos(\mathbf{z}_i, \mathbf{z}_j)}{2}\right], \tag{4}$$

where $\lambda \in (0, 1]$ balances relevance and redundancy. The second term penalizes similarity to the most similar selected reference and is defined as zero when $\mathcal{S}$ is empty. The first reference is therefore selected by relevance, while later selections seek distinct reference patterns in the relational representation. Selection stops when $|\mathcal{S}| = K$, and the resulting reference set $\mathcal{S}_q = \mathcal{S}$ is passed to the anomaly detector f_θ. Note that the selection procedure applies during training and inference.

3.3 Reference-Conditioned Anomaly Detection

Our selector is not tied to a specific anomaly detector. For a query $\boldsymbol{q}$, Π_K constructs a query-dependent reference set $\mathcal{S}_q$ from the corresponding normal pool. A reference-conditioned detector then predicts its anomaly map as $\widehat{\boldsymbol{y}}_q = f_\theta(\boldsymbol{q}, \mathcal{S}_q)$. Since Π_K operates independently of the detector parameters, it supports two integration modes. For a trainable detector, the selected references can guide detector training. For an already trained or training-free detector, they can replace its original reference input at inference without updating the detector. We instantiate the first mode with ResAD++ (Yao et al., 2025) and the second with UniVAD (Gu et al., 2025).

Selector-guided detector training. We adopt the detection architecture of ResAD++, including its frozen pretrained WideResNet-50-2 backbone. For each labeled source query $(\boldsymbol{q}, \boldsymbol{y}_q) \in \mathcal{D}_s$, we construct $\mathcal{S}_q = \Pi_K(\boldsymbol{q}, \mathcal{R}_c^s \setminus \{\boldsymbol{q}\}; E_\psi)$, where $\mathcal{R}_c^s$ is the normal reference pool of the corresponding source category. The detector is optimized by

$$\min_\theta \mathcal{L}_\theta = \frac{1}{|\mathcal{D}_s|} \sum_{(\boldsymbol{q}, \boldsymbol{y}_q) \in \mathcal{D}_s} \left[\mathcal{L}_{\text{focal}}\big(f_\theta(\boldsymbol{q}, \mathcal{S}_q), \boldsymbol{y}_q\big) + \mathcal{L}_{\text{reg}}\big(\theta; \boldsymbol{q}, \mathcal{S}_q, \boldsymbol{y}_q\big) \right]. \tag{5}$$

Thus, the loss for each query is computed using its own selected reference set. Here, $\mathcal{L}_{\text{focal}}$ is the pixel-wise focal loss (Lin et al., 2017). The term $\mathcal{L}_{\text{reg}}$ contains the auxiliary dictionary, residual mapping, and density learning objectives inherited from ResAD++, as detailed in Appendix C. The backbone and E_ψ remain fixed, and only θ is updated.

Plug-in use with a fixed detector. For an already trained or training-free detector $f_{\bar{\theta}}$, its parameters $\bar{\theta}$ remain fixed. For each target query $\boldsymbol{q}$, we select $\mathcal{S}_q = \Pi_K(\boldsymbol{q}, \mathcal{R}^t; E_\psi)$ and obtain $\widehat{\boldsymbol{y}}_q = f_{\bar{\theta}}(\boldsymbol{q}, \mathcal{S}_q)$, where $\mathcal{R}^t$ is the normal reference pool of the corresponding target category. For UniVAD, $\mathcal{S}_q$ replaces its original normal reference set, while its detection modules, scoring, and post-processing procedures remain unchanged.

3.4 Generalization Analysis

We analyze generalization from historical tasks to an unseen task. A task is denoted by $T = (\mathcal{R}, \{(\boldsymbol{q}_i, \boldsymbol{y}_i)\}_{i=1}^n)$, where $\mathcal{R}$ is a normal reference pool of size N and $\boldsymbol{y}_i \in [0, 1]^{H \times W}$ is the query label. Let $T_1, \ldots, T_m$ denote the historical tasks, and let $\mathfrak{P}$ denote the underlying task distribution. Let $\Pi_{\psi,K}(\boldsymbol{q}, \mathcal{R})$ denote the ordered sequence returned by Top-M retrieval followed by K MMR steps. Although the detector can be trained by focal loss in (5), other loss functions are also applicable for measuring the error or risk, such as mean square error, mean absolute error, and hinge loss. For convenience, we use the pixel-averaged ℓ_1 loss for the theoretical analysis, denoted as $\ell_{\psi,\theta}(\boldsymbol{q}, \mathcal{R}, \boldsymbol{y}) = \frac{1}{HW} \left\| f_\theta\big(\boldsymbol{q}, \Pi_{\psi,K}(\boldsymbol{q}, \mathcal{R})\big) - \boldsymbol{y} \right\|_1$, where detector outputs lie in $[0, 1]^{H \times W}$. Define $L_{\psi,\theta}(T) = \frac{1}{n} \sum_{i=1}^n \ell_{\psi,\theta}(\boldsymbol{q}_i, \mathcal{R}, \boldsymbol{y}_i)$. Then, the empirical risk $\widehat{R}(\psi, \theta) = \frac{1}{m} \sum_{t=1}^m L_{\psi,\theta}(T_t)$ and the true risk $R(\psi, \theta) = \mathbb{E}_{T \sim \mathfrak{P}} L_{\psi,\theta}(T)$.

Theorem 1 (Generalization with reference selection)**.** *Let the spectrum norm adjacency in GIN* $\|I + A_{\mathcal{E}}\|_2 \le a$, $\max_l \|W_{\psi,l}\|_2 \le c, \max_l \|W_{\theta,l}\|_2 \le c$, $\max_l \|W_{\psi,l}\|_F \le b$, *and* $\max_l \|W_{\theta,l}\|_F \le b$ *, for constant* $a, b, c > 0$*. Assume a common input bound* $\max\left\{\sup_x \|v_\psi(x)\|_F, \sup_{q,\mathcal{S}} \|v_\theta(q, \mathcal{S})\|_F\right\} \le c_v$*., where* v_ψ *and* v_θ *are the fixed input feature extraction. Assume* $\inf_{\psi,\boldsymbol{x}} \|E_\psi(\boldsymbol{x})\|_2 \ge \nu > 0$*, and all activation function are* 1*-Lipschitz. Let* $u > 0$ *be the covering radius of the detector class under the uniform pixel-averaged* L_1 *output distance. Then, for any fixed* $\gamma, u > 0$ *and* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$ *over* $T_1, \ldots, T_m$*, every admissible* (ψ, θ) *satisfies*

$$R(\psi, \theta) \le \widehat{R}(\psi, \theta) + \widehat{p}_\psi(\gamma) + 2u + \sqrt{\frac{\mathcal{C}(\gamma, u) + \log(1/\delta)}{2m}}, \tag{6}$$

where $\widehat{p}_\psi(\gamma)$ is the empirical fraction of queries whose minimum selection margin is at most γ, detailed in Appendix H, and $\mathcal{C}(\gamma, u) = L_\psi d_\psi^2 \log\left(1 + \frac{8b\sqrt{L_\psi}\kappa_\psi}{\gamma}\right) + L_\theta d_\theta^2 \log\left(1 + \frac{2b\sqrt{L_\theta}\kappa_\theta}{u}\right)$ *with constants* $\kappa_\psi = \frac{2ac_v\sqrt{L_\psi}}{\nu}(1+ac)^{L_\psi - 1}, \kappa_\theta = \frac{ac_v\sqrt{L_\theta}}{\sqrt{HW}}(1+ac)^{L_\theta - 1}$.

The proof is provided in Appendix H. The bounded-loss and activation assumptions are readily satisfied by predictions in $[0,1]^{H\times W}$ and ReLU, respectively. The theorem provides two implications.

- The term $\widehat{p}_\psi(\gamma)$ links generalization to reference selection stability. Larger retrieval and MMR margins preserve the selected references under encoder perturbations and reduce this penalty at a fixed γ. The bound therefore favors representations that achieve both low empirical detection error and stable reference choices.
- For fixed complexity, the concentration term decreases as $m^{-1/2}$, supporting training on more independent historical tasks. Greater network depth, width, or amplification increases the complexity bound. This motivates controlling model capacity and graph aggregation amplification while learning from historical tasks.

4 Experiments

4.1 Experimental Setup

Datasets. We use MVTec AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MPDD (Jezek et al., 2021), MVTec LOCO (Bergmann et al., 2022a), and MVTec 3D (Bergmann et al., 2022b) for industrial anomaly detection. Only RGB images are used from MVTec 3D. For source-trained models, we train on MVTec AD and test on VisA, or train on VisA and test on the remaining datasets. Normal training images from each target category form its reference pool $\mathcal{R}$; target test images are used for evaluation.

Compared Methods. We compare with representative anomaly detection methods with publicly available implementations, grouped into four categories. Conventional feature-based methods include **SPADE** (Cohen & Hoshen, 2020), **PaDiM** (Defard et al., 2021), and **PatchCore** (Roth et al., 2022). Learned reference-conditioned detectors include **RegAD** (Huang et al., 2022), **ResAD++** (Yao et al., 2025), **D^{2}4FAD** (Dong et al., 2026), and **InCTRL** (Zhu & Pang, 2024). **UniVAD** (Gu et al., 2025) represents training-free detection with vision foundation models, while **AnomalyCLIP** (Zhou et al., 2024) provides a zero-shot baseline. All reference-based methods use eight normal reference images, matching the reference budget of our detector. For InCTRL, we report the 8-shot results reported by ResAD++ (Yao et al., 2025). For the remaining methods, we use the authors' released implementations to rerun the evaluations.

To evaluate our selector's effectiveness across detectors, we compare GRef with four reference selection baselines on PatchCore, RegAD, ResAD++, and UniVAD, keeping detector parameters fixed. **Fixed** shares one reference set across queries within each category, **Random** samples independently for each query, **Similarity** retrieves by cosine similarity between globally pooled WideResNet-50-2 features (Zagoruyko & Komodakis, 2016), and **CPR** uses its global retrieval module (Li et al., 2024). All strategies use the same normal pool and eight references per query. To assess the additional benefit of selector-guided training, we also train RegAD and ResAD++ on source data using the frozen GRef selector. These variants, denoted **RegAD+GRef**† and **ResAD++ +GRef**†, use GRef during both training and inference and are compared with their inference-only counterparts.

Implementation and Evaluation Metrics. We use frozen WideResNet-50-2 (Zagoruyko & Komodakis, 2016) features. The relational encoder E_ψ is pretrained on source normal images and remains frozen during detector training and inference. By default, $M = 32$, $K = 8$, and $\lambda = 0.7$. Implementation details appear in Appendix D. Following ResAD++ (Yao et al., 2025), we use Image AUROC, Pixel AUROC, and AUPRO. Image AUROC measures image-level anomaly detection and is reported in the main text. Pixel AUROC and AUPRO measure anomaly localization and are reported in Appendix G. AUPRO is integrated up to a false-positive rate of 0.3.

Table 1: Cross-dataset Image AUROC (%). Our methods are evaluated over three runs. Mean denotes the average Image AUROC, and std stands for the sample standard deviation across runs. GRef (Ours) denotes ResAD++ trained with our reference selection method, while UniVAD + GRef uses our GRef selector directly at inference. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
SPADE	$78.67_{\pm 0.49}$	$74.80_{\pm 0.85}$	$85.15_{\pm 1.04}$	$66.34_{\pm 3.41}$	$64.24_{\pm 0.37}$	$60.27_{\pm 0.96}$	$71.58_{\pm 0.53}$
PaDiM	$83.01_{\pm 1.37}$	$72.61_{\pm 0.96}$	$92.65_{\pm 1.62}$	$60.53_{\pm 3.28}$	$56.38_{\pm 0.20}$	$63.23_{\pm 1.09}$	$71.40_{\pm 0.55}$
PatchCore	$92.95_{\pm 0.69}$	$87.56_{\pm 0.66}$	$92.00_{\pm 0.50}$	$79.17_{\pm 2.11}$	$64.26_{\pm 0.54}$	$71.89_{\pm 1.35}$	$81.31_{\pm 0.24}$
RegAD	$86.73_{\pm 0.22}$	$78.47_{\pm 1.46}$	$92.09_{\pm 1.07}$	$67.98_{\pm 2.11}$	$65.69_{\pm 1.61}$	$64.23_{\pm 0.81}$	$75.86_{\pm 0.77}$
ResAD++	$92.24_{\pm 0.75}$	$89.08_{\pm 0.79}$	$93.09_{\pm 0.45}$	$75.04_{\pm 1.17}$	$68.06_{\pm 0.32}$	$73.61_{\pm 1.81}$	$81.85_{\pm 0.17}$
$D^2$4FAD	$78.64_{\pm 0.30}$	$71.83_{\pm 0.69}$	$83.49_{\pm 0.74}$	$68.99_{\pm 2.14}$	$66.88_{\pm 0.46}$	$57.64_{\pm 1.67}$	$71.24_{\pm 0.62}$
AnomalyCLIP	$74.59_{\pm 0.92}$	$80.54_{\pm 0.26}$	$74.26_{\pm 1.41}$	$68.01_{\pm 0.96}$	$50.64_{\pm 0.46}$	$57.52_{\pm 0.90}$	$67.59_{\pm 0.59}$
InCTRL	$93.23_{\pm 0.63}$	$89.77_{\pm 0.51}$	$79.74_{\pm 0.65}$	$85.15_{\pm 0.93}$	$67.00_{\pm 0.82}$	$78.07_{\pm 0.57}$	$82.16_{\pm 0.44}$
GRef (Ours)	$\mathbf{97.45}_{\pm 0.10}$	$\mathbf{92.77}_{\pm 0.10}$	$\mathbf{94.17}_{\pm 0.00}$	$\mathbf{93.06}_{\pm 0.17}$	$\mathbf{71.20}_{\pm 0.17}$	$\mathbf{78.99}_{\pm 0.20}$	$\mathbf{87.94}_{\pm 0.00}$
UniVAD	$98.44_{\pm 0.39}$	$96.30_{\pm 0.26}$	$\mathbf{96.80}_{\pm 0.81}$	$90.77_{\pm 0.57}$	$77.85_{\pm 0.44}$	$90.55_{\pm 0.32}$	$91.79_{\pm 0.35}$
UniVAD + GRef (Ours)	$\mathbf{99.31}_{\pm 0.00}$	$\mathbf{96.40}_{\pm 0.14}$	$96.44_{\pm 0.00}$	$\mathbf{97.72}_{\pm 0.00}$	$\mathbf{80.35}_{\pm 0.14}$	$\mathbf{91.97}_{\pm 0.00}$	$\mathbf{93.70}_{\pm 0.01}$

Table 2: Cross-dataset Image AUROC (%) with eight reference images. Results report mean $\pm$ standard deviation over three runs. Avg. is the unweighted mean across the six datasets. † denotes detectors trained with GRef-selected references. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
PatchCore + Fixed	$92.95_{\pm 0.69}$	$87.56_{\pm 0.66}$	$92.00_{\pm 0.50}$	$79.17_{\pm 2.11}$	$64.26_{\pm 0.54}$	$71.89_{\pm 1.35}$	$81.31_{\pm 0.24}$
PatchCore + Random	$92.69_{\pm 0.77}$	$87.41_{\pm 0.51}$	$91.65_{\pm 0.35}$	$76.00_{\pm 2.17}$	$65.19_{\pm 0.48}$	$71.55_{\pm 0.33}$	$80.75_{\pm 0.65}$
PatchCore + Similarity	$95.87_{\pm 0.10}$	$88.24_{\pm 0.08}$	$90.21_{\pm 0.33}$	$90.41_{\pm 0.32}$	$65.98_{\pm 0.69}$	$74.68_{\pm 0.55}$	$84.23_{\pm 0.09}$
PatchCore + CPR	$96.95_{\pm 0.37}$	$89.97_{\pm 0.23}$	$\mathbf{93.00}_{\pm 0.48}$	$91.55_{\pm 0.69}$	$65.49_{\pm 0.47}$	$76.74_{\pm 0.69}$	$85.62_{\pm 0.18}$
PatchCore + GRef	$\mathbf{98.17}_{\pm 0.10}$	$\mathbf{91.04}_{\pm 0.41}$	$92.46_{\pm 0.77}$	$\mathbf{93.79}_{\pm 0.24}$	$\mathbf{69.16}_{\pm 0.37}$	$\mathbf{77.91}_{\pm 0.46}$	$\mathbf{87.09}_{\pm 0.24}$
RegAD + Fixed	$86.73_{\pm 0.22}$	$78.47_{\pm 1.46}$	$92.09_{\pm 1.07}$	$67.98_{\pm 2.11}$	$65.69_{\pm 1.61}$	$64.23_{\pm 0.81}$	$75.86_{\pm 0.77}$
RegAD + Random	$86.83_{\pm 1.84}$	$77.84_{\pm 0.35}$	$92.05_{\pm 0.40}$	$68.45_{\pm 0.88}$	$64.69_{\pm 0.97}$	$64.03_{\pm 0.69}$	$75.65_{\pm 0.68}$
RegAD + Similarity	$90.37_{\pm 0.94}$	$80.38_{\pm 1.06}$	$90.50_{\pm 1.04}$	$80.24_{\pm 0.70}$	$68.49_{\pm 0.65}$	$69.90_{\pm 1.03}$	$79.98_{\pm 0.54}$
RegAD + CPR	$90.47_{\pm 1.25}$	$82.50_{\pm 0.38}$	$92.40_{\pm 1.03}$	$\mathbf{83.11}_{\pm 2.07}$	$69.93_{\pm 1.24}$	$70.94_{\pm 0.98}$	$81.56_{\pm 1.07}$
RegAD + GRef	$\mathbf{93.00}_{\pm 0.84}$	$84.37_{\pm 0.78}$	$\mathbf{92.62}_{\pm 0.53}$	$81.70_{\pm 0.26}$	$\mathbf{72.95}_{\pm 0.91}$	$71.93_{\pm 0.89}$	$\mathbf{82.76}_{\pm 0.33}$
RegAD + GRef†	$92.66_{\pm 0.49}$	$\mathbf{84.77}_{\pm 0.52}$	$92.01_{\pm 0.10}$	$81.84_{\pm 0.33}$	$71.15_{\pm 1.47}$	$\mathbf{72.12}_{\pm 1.59}$	$82.42_{\pm 0.47}$
ResAD++ + Fixed	$92.24_{\pm 0.75}$	$89.08_{\pm 0.79}$	$93.09_{\pm 0.45}$	$75.04_{\pm 1.17}$	$68.06_{\pm 0.32}$	$73.61_{\pm 1.81}$	$81.85_{\pm 0.17}$
ResAD++ + Random	$92.08_{\pm 0.88}$	$89.31_{\pm 0.22}$	$93.72_{\pm 0.81}$	$76.50_{\pm 1.39}$	$67.18_{\pm 0.89}$	$72.53_{\pm 1.41}$	$81.89_{\pm 0.26}$
ResAD++ + Similarity	$95.46_{\pm 0.04}$	$89.52_{\pm 0.38}$	$92.22_{\pm 0.05}$	$89.93_{\pm 0.15}$	$68.06_{\pm 0.15}$	$76.92_{\pm 0.08}$	$85.35_{\pm 0.08}$
ResAD++ + CPR	$95.67_{\pm 0.57}$	$90.09_{\pm 0.40}$	$93.74_{\pm 0.44}$	$92.70_{\pm 0.48}$	$67.98_{\pm 0.87}$	$78.46_{\pm 0.21}$	$86.44_{\pm 0.14}$
ResAD++ + GRef	$97.34_{\pm 0.00}$	$92.43_{\pm 0.36}$	$94.08_{\pm 0.14}$	$92.03_{\pm 0.14}$	$70.32_{\pm 0.14}$	$\mathbf{80.43}_{\pm 0.00}$	$87.77_{\pm 0.10}$
ResAD++ + GRef†	$\mathbf{97.45}_{\pm 0.10}$	$\mathbf{92.77}_{\pm 0.10}$	$\mathbf{94.17}_{\pm 0.00}$	$\mathbf{93.06}_{\pm 0.17}$	$\mathbf{71.20}_{\pm 0.17}$	$78.99_{\pm 0.20}$	$\mathbf{87.94}_{\pm 0.00}$
UniVAD + Fixed	$98.44_{\pm 0.39}$	$96.30_{\pm 0.26}$	$96.80_{\pm 0.81}$	$90.77_{\pm 0.57}$	$77.85_{\pm 0.44}$	$90.55_{\pm 0.32}$	$91.79_{\pm 0.35}$
UniVAD + Random	$97.77_{\pm 0.48}$	$96.05_{\pm 0.06}$	$95.38_{\pm 0.53}$	$91.49_{\pm 0.54}$	$77.49_{\pm 0.42}$	$89.98_{\pm 0.26}$	$91.36_{\pm 0.15}$
UniVAD + Similarity	$98.69_{\pm 0.07}$	$95.97_{\pm 0.11}$	$96.87_{\pm 0.52}$	$\mathbf{98.07}_{\pm 0.08}$	$79.55_{\pm 0.19}$	$\mathbf{92.53}_{\pm 0.13}$	$93.61_{\pm 0.10}$
UniVAD + CPR	$98.75_{\pm 0.26}$	$\mathbf{96.57}_{\pm 0.08}$	$\mathbf{96.95}_{\pm 0.82}$	$97.52_{\pm 0.23}$	$79.38_{\pm 0.35}$	$92.48_{\pm 0.16}$	$93.61_{\pm 0.16}$
UniVAD + GRef	$\mathbf{99.31}_{\pm 0.00}$	$96.40_{\pm 0.14}$	$96.44_{\pm 0.00}$	$97.72_{\pm 0.00}$	$\mathbf{80.35}_{\pm 0.14}$	$91.97_{\pm 0.00}$	$\mathbf{93.70}_{\pm 0.01}$

4.2 Cross-Dataset Comparison

Table 1 compares detection performance across six industrial datasets. With selector-guided training, GRef achieves an average Image AUROC of 87.94%, outperforming all compared learned detectors on every dataset. Integrating our selector into UniVAD at inference yields the best overall average of 93.70%. These results support our claim that reference selection improves detection on unseen products without target-side parameter updates. Table 2 compares selection strategies under the same normal pool and reference budget with detector parameters fixed. GRef achieves the highest average Image AUROC for all four detectors, outperforming random sampling and the similarity-based retrieval baselines. This supports considering both reference relevance and set composition and demonstrates compatibility with different detectors. The † variants show that selector-guided training provides a further benefit for ResAD++, but not for RegAD on average. Thus, effective inference-time selection does not require retraining the detector.

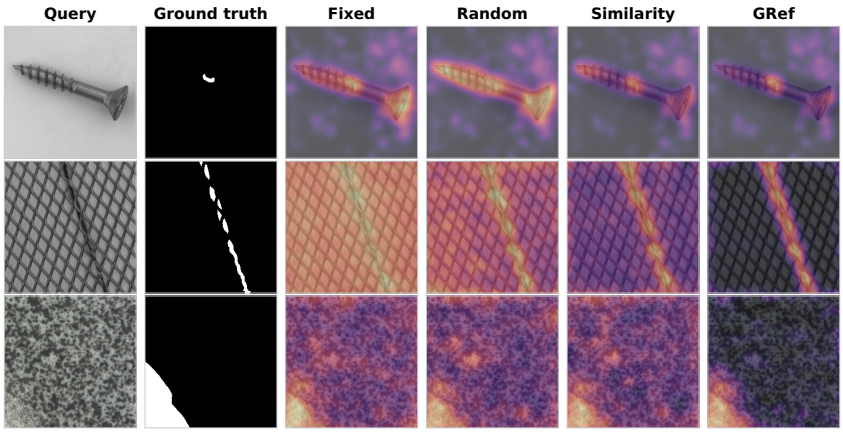

Figure 4: Qualitative comparison of reference selection strategies. Each row shows a query image, its ground-truth anomaly mask, and anomaly maps obtained with Fixed, Random, Similarity, and GRef reference selection, from left to right. White regions in the ground-truth masks indicate anomalies.

4.3 Ablation Study

Effect of the Reference Budget We vary the candidate budget M and reference budget K for MVTec AD-to-VisA transfer, keeping the normal reference pool, detector parameters, and $\lambda = 0.7$ fixed. Figure 3 reports Image AUROC for each (M, K) pair, with K on the horizontal axis and M on the vertical axis. Increasing K consistently improves Image AUROC across the evaluated candidate budgets, whereas increasing M does not yield consistent gains at a fixed K. Additional parameter studies are given in Appendix E.

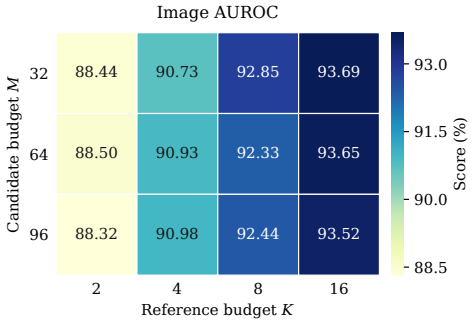

Figure 3: Effect of candidate budget M and reference budget K on MVTec AD-to-VisA transfer. Each cell reports Image AUROC (%).

Qualitative Comparison Figure 4 compares anomaly maps obtained with Fixed, Random, Similarity, and GRef reference selection for three query images. In these examples, Fixed and Random produce diffuse responses over normal regions. Similarity reduces some of these responses, while GRef yields more localized activation around the annotated defects. In particular, GRef suppresses responses on the intact grid pattern and reduces background activation in the bottom example. Inference speed and efficiency are analyzed in Appendix F.

5 Conclusion and Limitations

We presented GRef, a graph-guided framework for selecting normal references for anomaly detection on unseen products. By combining transferable relational representations with selection that balances relevance and redundancy, it supports both selector-guided detector training and plug-in inference without target-side parameter updates. Our generalization analysis accounts for discrete reference selection, and experiments across six industrial datasets demonstrate its effectiveness with different detectors. A limitation is that the selector is pretrained independently and remains frozen during detector training. Its retrieval representations may therefore not fully align with the detector's needs, potentially limiting generalization to unseen products. Future work will explore joint training of the selector and detector to better align reference selection with detection objectives.

AI USE STATEMENT

Generative AI tools were used for coding assistance, checking theoretical arguments, and minor language editing. The authors are responsible for the final manuscript, code, and results.

REFERENCES

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2(3):5, 2018.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.pdf.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 2022a. doi: 10.1007/s11263-022-01578-9. URL https://www.mvtec.com/research-teaching/datasets/mvtec-loco-ad.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, 2022b. URL https://www.scitepress.org/PublishedPapers/2022/108650/.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. URL https://arxiv.org/abs/2005.02357.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12664 of *Lecture Notes in Computer Science*, pp. 475–489. Springer, 2021. doi: 10.1007/978-3-030-68799-1_35. URL https://link.springer.com/chapter/10.1007/978-3-030-68799-1_35.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Le Dong, Qinzhong Tan, Chunlei Li, Jingliang Hu, Yilei Shi, Weisheng Dong, Xiao Xiang Zhu, and Lichao Mou. Dual distillation for few-shot anomaly detection. In *International Conference on Learning Representations*, 2026.

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17481–17490,

2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Fang_FastRecon_Few-shot_Industrial_Anomaly_Detection_via_Fast_Feature_Reconstruction_ICCV_2023_paper.html.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15194–15203, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gu_UniVAD_A_Training-free_Unified_Model_for_Few-shot_Visual_Anomaly_Detection_CVPR_2025_paper.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pp. 303–319, 2022. doi: 10.1007/978-3-031-20053-3_18. URL https://arxiv.org/abs/2207.07361.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 66–71, 2021. doi: 10.1109/ICUMT54235.2021.9631567. URL https://github.com/stepanje/MPDD.

Le Jiang, Yan Huang, Zhen Xu, Yong Xu, Hau-San Wong, and Si Wu. Defect cue-preserved structural feature refinement for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35607–35616, 2026.

Hanxi Li, Jianfei Hu, Bo Li, Hao Chen, Yongbin Zheng, and Chunhua Shen. Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval. *IEEE Transactions on Image Processing*, 33:5606–5621, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023. doi: 10.1109/CVPR52729.2023.01954. URL https://arxiv.org/abs/2303.15140.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *IEEE International Symposium on Industrial Electronics*, pp. 1–6, 2021. doi: 10.1109/ISIE45552.2021.9576231. URL https://arxiv.org/abs/2104.10036.

Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Xiaopei Zhang, Xingang Wang, Fei Shen, Zhengtao Zhang, Mukesh Prasad, and Guiguang Ding. DictAS: A framework for class-generalizable few-shot anomaly segmentation via dictionary lookup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20519–20528, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Qu_DictAS_A_Framework_for_Class-Generalizable_Few-Shot_Anomaly_Segmentation_via_Dictionary_ICCV_2025_paper.html.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition, 2026. doi: 10.1017/9781009490672.

Xiaolei Wang, Yuexin Wang, Tianhong Dai, Huihui Bai, Yao Zhao, and Jimin Xiao. Hunting normality from query sample via residual learning for generalist anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43093–43102, 2026.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.00826.

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. Resad: A simple framework for class generalizable anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125287–125311, 2024. doi: 10.52202/079017-3979.

Xincheng Yao, Chao Shi, Muming Zhao, Guangtao Zhai, and Chongyang Zhang. ResAD++: Towards class agnostic anomaly detection via residual feature learning, 2025. URL https://arxiv.org/abs/2509.23741.

Eunsun Yun, Heechul Lim, Haeyun Lee, Kang-Wook Chon, and Minjoong Jeong. QueCo: Query-conditioned consensus over unverified evidence for training-free zero-shot anomaly detection. *Knowledge-Based Systems*, pp. 116771, 2026. doi: 10.1016/j.knosys.2026.116771.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pp. 87.1–87.12, 2016. doi: 10.5244/C.30.87. URL https://bmva-archive.org.uk/bmvc/2016/papers/paper087/index.html.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.18961.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17826–17836, June 2024. doi: 10.1109/CVPR52733.2024.01688. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhu_Toward_Generalist_Anomaly_Detection_via_In-context_Residual_Learning_with_Few-shot_CVPR_2024_paper.html.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408, 2022. doi: 10.1007/978-3-031-20056-4_23. URL https://arxiv.org/abs/2207.14315.

A Relational Encoder

The frozen feature extractor ϕ produces the image feature map F_3. Adaptive average pooling over the $2^l \times 2^l$ grids in Section 3.1 converts this map into regional features at multiple spatial scales.

Each region $R_{l,r,c}$ forms a graph node. Bidirectional edges connect horizontal and vertical neighbors within each grid and parent–child regions across consecutive levels. GIN updates the regional features through these connections, with a residual connection to the original features V. The readout in Eq. (2) maps the updated features to the image representation $z(x) = E_\psi(x)$.

The pretraining objective in Eq. (3) combines representation consistency with relational supervision. Two independently sampled feature-dropout views are generated for each source normal image. The consistency term aligns their representations, while the relation term uses similarities between clean regional features to supervise similarities between perturbed representations. The dropout masks are resampled at each optimization step.

B Reference Selection

For query q, references $r_i \in \mathcal{R}$ are ranked by $g_i = (1 + \cos(z_q, z_i))/2$, where $z_q = E_\psi(q)$ and $z_i = E_\psi(r_i)$. The top M images form the candidate set C_q. The selection rule in Eq. (4) then builds S_q one reference at a time, balancing query relevance against similarity to the references already selected. The redundancy penalty is zero for the first reference, and selection stops after K distinct images have been chosen.

Reference representations can be computed once with the frozen encoder and reused across queries. The default settings are $M = 32$, $K = 8$, and $\lambda = 0.7$. Budget comparisons use the same normal reference pool.

Selection controls. Fixed uses the same eight normal images for every query in a category, whereas Random samples eight images separately for each query. Similarity ranks the full normal pool by cosine similarity between 2048-dimensional, globally pooled WideResNet features. CPR ranks the same pool using visual-word histograms constructed from the spatial descriptors in F_3. Both select eight references directly from the full pool, and their per-query reference lists are shared across detectors. relational encoding, candidate retrieval, and query-adaptive reference set selection.

C Anomaly Detectors

The detector predicts an anomaly map conditioned on the selected references, $\hat{y}_q = f_\theta(q, S_q)$. Our ResAD++ instantiation (Yao et al., 2025) matches query and reference features to form residual features, then applies feature-distribution matching, residual mapping, and density estimation. Its feature extractor remains frozen.

Detector objective. Equation (5) distinguishes the pixel-wise focal loss from the auxiliary detector objectives in $\mathcal{L}_{\text{reg}}$. For ResAD++, these objectives concern the normal residual-feature dictionary, residual-feature constraints, and distribution estimation.

Reference-based comparisons. SPADE (Cohen & Hoshen, 2020), PaDiM (Defard et al., 2021), and PatchCore (Roth et al., 2022) use pretrained features for multiscale matching, Gaussian feature modeling, and nearest-neighbor scoring, respectively. Their reference features or statistics are constructed from the eight supplied normal images. RegAD (Huang et al., 2022) learns feature registration, ResAD++ (Yao et al., 2025) models residual features, and UniVAD (Gu et al., 2025) performs component-aware matching. In the reference-selection comparisons, each detector retains its detection procedure while receiving the references selected by the corresponding strategy.

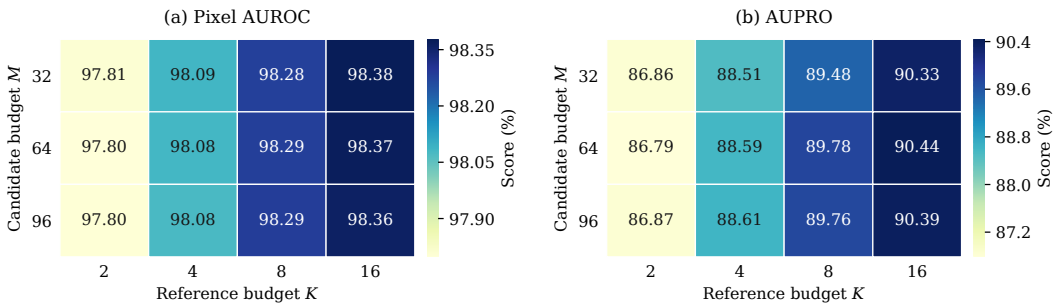

Figure 5: Localization performance across candidate and reference budgets on MVTec AD-to-VisA transfer. The two heatmaps report Pixel AUROC and AUPRO (%), respectively, with reference budget K on the horizontal axis and candidate budget M on the vertical axis. The relevance weight is fixed at $\lambda = 0.7$.

Table 3: Effect of the relevance weight on VisA-to-MVTec AD transfer. We fix $M = 32$ and $K = 8$. All metrics are in percent. Results are obtained from a separate single-run evaluation for each setting.

λ	Image AUROC	Pixel AUROC	AUPRO
0.5	97.58	97.48	92.59
0.7	97.62	97.53	92.64
0.9	97.50	97.48	92.63

D Implementation Details

Relational encoder pretraining. The relational encoder is pretrained on source normal images for 25 epochs using AdamW, with a learning rate of 10^{-4}, weight decay of 10^{-4}, and batch size of 64. The feature extractor is frozen throughout, and the relational encoder is frozen after pretraining.

Selector-guided detector training. The detector is trained with the selected normal references while the relational encoder remains frozen. The learned ResAD++ modules use Adam with a learning rate of 10^{-5} for 50 epochs; its feature extractor remains frozen. RegAD uses SGD with an initial learning rate of 10^{-4} and cosine decay for 50 epochs. Both use a batch size of 32 and eight references.

Input and hardware. GRef uses 224×224 input images. All experiments run on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs.

E Parameter Studies

Candidate and reference budgets. We vary the candidate budget M and reference budget K on MVTec AD-to-VisA transfer while fixing the detector weights, normal reference pool, and $\lambda = 0.7$. Figure 5 reports Pixel AUROC and AUPRO for the same (M, K) combinations as the image-level evaluation in Figure 3. Increasing K consistently improves both localization metrics across the evaluated candidate budgets, whereas the effect of increasing M is less consistent.

Relevance weight. Table 3 compares $\lambda \in \{0.5, 0.7, 0.9\}$ with $M = 32$ and $K = 8$. This parameter balances query relevance and reference redundancy.

F Inference Efficiency

Figure 6 compares inference speed, detection performance, and model size on MVTec AD. We measure inference latency on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU

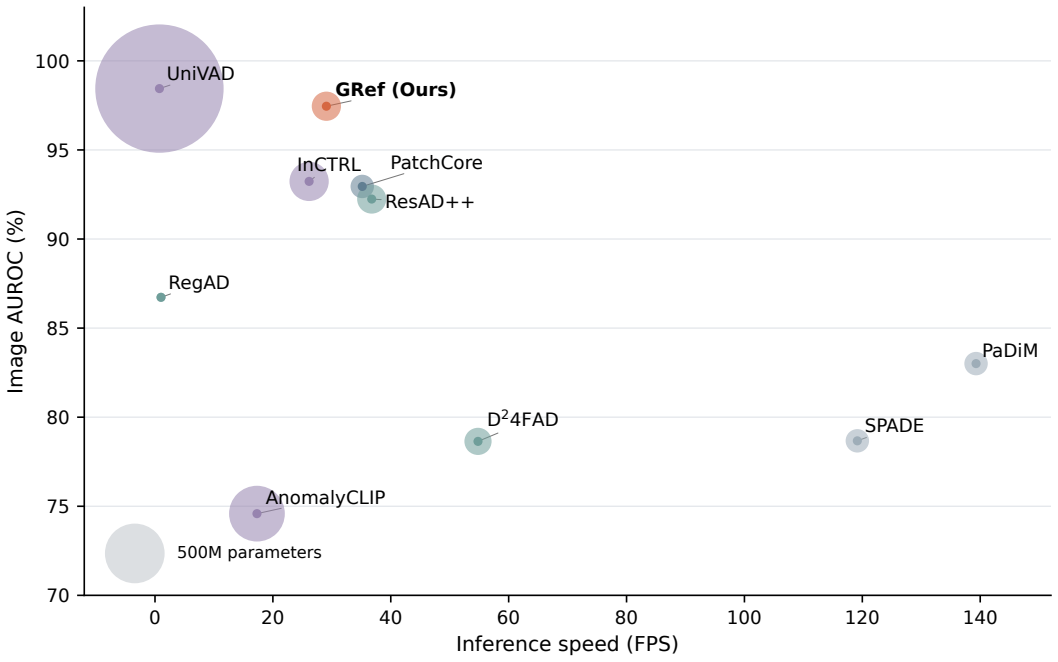

Figure 6: Inference speed and detection performance on MVTec AD. The horizontal axis reports inference speed in FPS, and the vertical axis reports Image AUROC (%). Bubble area represents the total number of model parameters, including frozen parameters; reference caches are excluded. GRef timing includes reference selection with $M = 32$ and $K = 8$.

with batch size one, using each method's native input resolution. Inference speed is measured on a shared subset of 75 test images from all 15 MVTec AD categories, with five images per category. After warm-up, each image is evaluated three times, with GPU synchronization before and after timing, yielding 225 timed inference calls per method. Image AUROC is taken from the full-dataset evaluation over three runs.

Timing excludes data loading, input preprocessing and transfer, model initialization, offline reference preparation, and metric computation. It includes query feature extraction, matching, and postprocessing. For GRef, it also includes candidate retrieval and the selection of eight references from 32 candidates, with normal-reference representations cached in advance. FPS is computed as the reciprocal of mean per-image latency.

GRef achieves 29.07 FPS with an Image AUROC of 97.45%. Compared with ResAD++ at 36.76 FPS and 92.24% Image AUROC, it improves detection performance by 5.21 percentage points while increasing mean latency from 27.20 to 34.39 ms per image. UniVAD achieves a higher Image AUROC of 98.44% but runs at 0.75 FPS under the same timing protocol. These measurements characterize the evaluated local implementations, including their caching and GPU matching optimizations.

G Additional Detection Results

Tables 4 and 5 report localization performance for the main comparison. Tables 6 and 7 compare reference-selection strategies using the same two metrics.

H Proof of Theorem 1

To prove Theorem 1, we first relate parameter perturbations to changes in the encoder and detector outputs. We then show that sufficiently large retrieval and MMR margins preserve the selected reference sequence. These two results let us construct finite model representatives and compare their

Table 4: Cross-dataset Pixel AUROC (%). Our methods are evaluated over three runs. Mean denotes the average Pixel AUROC, and std stands for the sample standard deviation across runs. GRef (Ours) denotes ResAD++ trained with our reference selection method, while UniVAD + GRef uses our GRef selector directly at inference. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
SPADE	$94.68_{\pm 0.32}$	$96.17_{\pm 0.32}$	$97.33_{\pm 0.10}$	$96.71_{\pm 0.00}$	$63.18_{\pm 0.22}$	$96.59_{\pm 0.00}$	$90.78_{\pm 0.00}$
PaDiM	$93.64_{\pm 0.24}$	$93.02_{\pm 0.48}$	$96.71_{\pm 0.00}$	$92.32_{\pm 0.14}$	$65.16_{\pm 0.75}$	$94.97_{\pm 0.00}$	$89.30_{\pm 0.14}$
PatchCore	$95.49_{\pm 0.37}$	$96.99_{\pm 0.26}$	$96.34_{\pm 0.10}$	$96.73_{\pm 0.10}$	$68.68_{\pm 0.24}$	$96.71_{\pm 0.00}$	$91.82_{\pm 0.00}$
RegAD	$96.46_{\pm 0.33}$	$96.29_{\pm 0.10}$	$97.48_{\pm 0.32}$	$95.09_{\pm 0.77}$	$71.15_{\pm 0.70}$	$96.95_{\pm 0.17}$	$92.24_{\pm 0.30}$
ResAD++	$95.78_{\pm 0.40}$	$97.07_{\pm 0.10}$	$97.54_{\pm 0.10}$	$97.80_{\pm 0.00}$	$66.91_{\pm 0.10}$	$97.42_{\pm 0.00}$	$92.09_{\pm 0.01}$
$D^2$4FAD	$79.90_{\pm 0.22}$	$75.43_{\pm 2.94}$	$85.65_{\pm 0.58}$	$74.45_{\pm 0.73}$	$\mathbf{74.18}_{\pm 0.86}$	$74.68_{\pm 1.37}$	$77.38_{\pm 0.71}$
AnomalyCLIP	$78.06_{\pm 1.14}$	$94.17_{\pm 0.14}$	$66.36_{\pm 1.00}$	$75.38_{\pm 1.50}$	$55.08_{\pm 0.67}$	$84.31_{\pm 0.44}$	$75.56_{\pm 0.22}$
GRef (Ours)	$\mathbf{97.50}_{\pm 0.00}$	$\mathbf{98.27}_{\pm 0.00}$	$\mathbf{97.88}_{\pm 0.00}$	$\mathbf{98.82}_{\pm 0.00}$	$68.23_{\pm 0.10}$	$\mathbf{97.80}_{\pm 0.00}$	$\mathbf{93.08}_{\pm 0.01}$
UniVAD	$96.67_{\pm 0.14}$	$98.53_{\pm 0.00}$	$97.87_{\pm 0.00}$	$98.46_{\pm 0.00}$	$\mathbf{74.16}_{\pm 0.00}$	$98.61_{\pm 0.00}$	$94.05_{\pm 0.02}$
UniVAD + GRef (Ours)	$\mathbf{97.16}_{\pm 0.00}$	$\mathbf{98.66}_{\pm 0.00}$	$\mathbf{98.02}_{\pm 0.00}$	$\mathbf{99.04}_{\pm 0.00}$	$\mathbf{74.16}_{\pm 0.10}$	$\mathbf{98.68}_{\pm 0.00}$	$\mathbf{94.29}_{\pm 0.01}$

Table 5: Cross-dataset AUPRO (%). Our methods are evaluated over three runs. Mean denotes the average AUPRO, and std stands for the sample standard deviation across runs. GRef (Ours) denotes ResAD++ trained with our reference selection method, while UniVAD + GRef uses our GRef selector directly at inference. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
SPADE	$86.38_{\pm 0.55}$	$80.96_{\pm 0.58}$	$80.32_{\pm 0.33}$	$88.25_{\pm 0.49}$	$61.59_{\pm 0.24}$	$89.45_{\pm 0.20}$	$81.16_{\pm 0.10}$
PaDiM	$82.95_{\pm 0.30}$	$65.03_{\pm 1.09}$	$76.47_{\pm 0.74}$	$75.02_{\pm 1.35}$	$58.55_{\pm 0.37}$	$84.06_{\pm 0.17}$	$73.68_{\pm 0.54}$
PatchCore	$84.96_{\pm 0.76}$	$82.37_{\pm 0.95}$	$71.48_{\pm 0.40}$	$87.91_{\pm 1.45}$	$61.01_{\pm 0.32}$	$88.20_{\pm 0.26}$	$79.32_{\pm 0.39}$
RegAD	$87.65_{\pm 0.96}$	$78.48_{\pm 1.24}$	$76.46_{\pm 1.72}$	$82.81_{\pm 2.63}$	$\mathbf{67.75}_{\pm 1.07}$	$89.73_{\pm 0.57}$	$80.48_{\pm 0.80}$
ResAD++	$88.86_{\pm 0.83}$	$86.36_{\pm 0.41}$	$78.55_{\pm 0.28}$	$91.88_{\pm 0.26}$	$59.29_{\pm 0.00}$	$91.16_{\pm 0.22}$	$82.68_{\pm 0.00}$
$D^2$4FAD	$32.70_{\pm 1.14}$	$20.32_{\pm 1.19}$	$41.59_{\pm 3.06}$	$32.86_{\pm 0.33}$	$32.01_{\pm 1.08}$	$31.87_{\pm 1.07}$	$31.89_{\pm 0.32}$
AnomalyCLIP	$51.93_{\pm 2.08}$	$86.86_{\pm 0.52}$	$22.79_{\pm 0.59}$	$45.79_{\pm 2.53}$	$33.95_{\pm 2.02}$	$50.33_{\pm 0.96}$	$48.61_{\pm 0.62}$
GRef (Ours)	$\mathbf{92.58}_{\pm 0.00}$	$\mathbf{89.46}_{\pm 0.10}$	$\mathbf{81.12}_{\pm 0.00}$	$\mathbf{94.61}_{\pm 0.00}$	$61.85_{\pm 0.17}$	$\mathbf{92.41}_{\pm 0.10}$	$\mathbf{85.34}_{\pm 0.03}$
UniVAD	$93.41_{\pm 0.17}$	$94.92_{\pm 0.30}$	$82.69_{\pm 0.86}$	$96.35_{\pm 0.00}$	$70.45_{\pm 0.20}$	$95.86_{\pm 0.00}$	$88.95_{\pm 0.10}$
UniVAD + GRef (Ours)	$\mathbf{94.16}_{\pm 0.00}$	$\mathbf{95.30}_{\pm 0.10}$	$\mathbf{83.07}_{\pm 0.00}$	$\mathbf{97.49}_{\pm 0.00}$	$\mathbf{71.84}_{\pm 0.10}$	$\mathbf{96.13}_{\pm 0.00}$	$\mathbf{89.66}_{\pm 0.02}$

losses with those of the original models. Finally, a bounded loss envelope handles the remaining small-margin inputs, and concentration across the m independent tasks yields the uniform generalization bound. We establish the loss properties and parameter conventions before following these steps. Additional interpretations are collected after the proof.

H.1 Setup and Loss Properties

For a realization T_t of a historical task, write

$$T_t = (\mathcal{R}_t, \{(q_{t,i}, y_{t,i})\}_{i=1}^{n_t}), \qquad n_t = n_{T_t}.$$

The sampling assumption is $T_1, \dots, T_m \overset{\text{iid}}{\sim} \mathfrak{P}$. Each task has finitely many queries, with $n_t \geq 1$, and the test task is an independent draw from $\mathfrak{P}$. The normal pool and queries may have arbitrary dependence within a task. In particular, multiple queries may share the same pool. No product factorization of the query and reference distributions is required. All feature maps and network outputs are measurable on the admissible input domain. References have fixed indices, and ties are resolved by a deterministic index rule.

For a prediction $\widehat{y}$ and annotation y, define

$$d_{\text{pix}}(\widehat{y}, y) = \frac{1}{HW} \|\widehat{y} - y\|_1. \tag{7}$$

For $h = (\psi, \theta)$ and $e = (q, \mathcal{R})$, the analysis loss is

$$\ell_h(e, y) = d_{\text{pix}}\big(f_\theta(q, \Pi_{\psi,K}(e)), y\big).$$

Table 6: Cross-dataset Pixel AUROC (%) with eight reference images. Results report mean ± standard deviation over three runs. Avg. is the unweighted mean across the six datasets. † denotes detectors trained with GRef-selected references. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
PatchCore + Fixed	$95.49_{\pm 0.37}$	$96.99_{\pm 0.26}$	$96.34_{\pm 0.10}$	$96.73_{\pm 0.10}$	$68.68_{\pm 0.24}$	$96.71_{\pm 0.00}$	$91.82_{\pm 0.00}$
PatchCore + Random	$95.63_{\pm 0.17}$	$96.94_{\pm 0.00}$	$96.34_{\pm 0.00}$	$96.57_{\pm 0.14}$	$69.07_{\pm 0.00}$	$96.55_{\pm 0.00}$	$91.85_{\pm 0.00}$
PatchCore + Similarity	$97.69_{\pm 0.01}$	$97.30_{\pm 0.03}$	$\mathbf{96.86}_{\pm 0.02}$	$97.62_{\pm 0.08}$	$70.21_{\pm 0.18}$	$97.46_{\pm 0.06}$	$92.86_{\pm 0.03}$
PatchCore + CPR	$97.51_{\pm 0.12}$	$97.32_{\pm 0.08}$	$96.14_{\pm 0.07}$	$97.91_{\pm 0.07}$	$70.35_{\pm 0.25}$	$97.45_{\pm 0.03}$	$92.78_{\pm 0.05}$
PatchCore + GRef	$\mathbf{97.90}_{\pm 0.00}$	$\mathbf{97.67}_{\pm 0.00}$	$96.83_{\pm 0.00}$	$\mathbf{98.06}_{\pm 0.00}$	$\mathbf{70.92}_{\pm 0.10}$	$\mathbf{97.77}_{\pm 0.00}$	$\mathbf{93.19}_{\pm 0.01}$
RegAD + Fixed	$96.46_{\pm 0.33}$	$96.29_{\pm 0.10}$	$97.48_{\pm 0.32}$	$95.09_{\pm 0.77}$	$71.15_{\pm 0.70}$	$96.95_{\pm 0.17}$	$92.24_{\pm 0.30}$
RegAD + Random	$96.48_{\pm 0.35}$	$96.27_{\pm 0.24}$	$97.49_{\pm 0.14}$	$95.13_{\pm 0.65}$	$71.19_{\pm 0.73}$	$96.94_{\pm 0.17}$	$92.25_{\pm 0.28}$
RegAD + Similarity	$96.68_{\pm 0.17}$	$96.88_{\pm 0.14}$	$97.59_{\pm 0.16}$	$96.35_{\pm 0.47}$	$71.74_{\pm 0.91}$	$97.33_{\pm 0.17}$	$92.76_{\pm 0.26}$
RegAD + CPR	$97.24_{\pm 0.18}$	$97.01_{\pm 0.13}$	$97.20_{\pm 0.26}$	$96.49_{\pm 0.47}$	$72.09_{\pm 1.03}$	$97.60_{\pm 0.16}$	$92.94_{\pm 0.31}$
RegAD + GRef	$\mathbf{97.66}_{\pm 0.14}$	$\mathbf{97.45}_{\pm 0.14}$	$\mathbf{97.84}_{\pm 0.17}$	$96.81_{\pm 0.47}$	$\mathbf{72.20}_{\pm 0.81}$	$\mathbf{97.83}_{\pm 0.17}$	$\mathbf{93.30}_{\pm 0.24}$
RegAD + GRef†	$97.52_{\pm 0.10}$	$97.29_{\pm 0.55}$	$97.79_{\pm 0.00}$	$\mathbf{97.21}_{\pm 0.14}$	$70.79_{\pm 1.16}$	$97.81_{\pm 0.14}$	$93.07_{\pm 0.14}$
ResAD++ + Fixed	$95.78_{\pm 0.40}$	$97.07_{\pm 0.10}$	$97.54_{\pm 0.10}$	$97.80_{\pm 0.00}$	$66.91_{\pm 0.10}$	$97.42_{\pm 0.00}$	$92.09_{\pm 0.01}$
ResAD++ + Random	$95.89_{\pm 0.00}$	$97.11_{\pm 0.10}$	$97.53_{\pm 0.10}$	$97.79_{\pm 0.00}$	$66.82_{\pm 0.20}$	$97.38_{\pm 0.10}$	$92.09_{\pm 0.00}$
ResAD++ + Similarity	$96.07_{\pm 0.01}$	$96.97_{\pm 0.04}$	$97.64_{\pm 0.01}$	$98.71_{\pm 0.01}$	$67.16_{\pm 0.00}$	$97.88_{\pm 0.02}$	$92.41_{\pm 0.01}$
ResAD++ + CPR	$96.96_{\pm 0.05}$	$97.35_{\pm 0.05}$	$97.22_{\pm 0.16}$	$98.58_{\pm 0.01}$	$67.62_{\pm 0.17}$	$97.82_{\pm 0.06}$	$92.59_{\pm 0.05}$
ResAD++ + GRef	$97.35_{\pm 0.00}$	$97.72_{\pm 0.01}$	$\mathbf{97.96}_{\pm 0.00}$	$\mathbf{98.84}_{\pm 0.00}$	$67.85_{\pm 0.00}$	$\mathbf{98.01}_{\pm 0.00}$	$92.95_{\pm 0.01}$
ResAD++ + GRef†	$\mathbf{97.50}_{\pm 0.00}$	$\mathbf{98.27}_{\pm 0.00}$	$97.88_{\pm 0.00}$	$98.82_{\pm 0.00}$	$\mathbf{68.23}_{\pm 0.10}$	$97.80_{\pm 0.00}$	$\mathbf{93.08}_{\pm 0.01}$
UniVAD + Fixed	$96.67_{\pm 0.14}$	$98.53_{\pm 0.00}$	$97.87_{\pm 0.00}$	$98.46_{\pm 0.00}$	$74.16_{\pm 0.00}$	$98.61_{\pm 0.00}$	$94.05_{\pm 0.02}$
UniVAD + Random	$96.78_{\pm 0.18}$	$98.52_{\pm 0.09}$	$97.87_{\pm 0.09}$	$98.51_{\pm 0.13}$	$74.15_{\pm 0.21}$	$98.60_{\pm 0.11}$	$94.07_{\pm 0.06}$
UniVAD + Similarity	$97.14_{\pm 0.01}$	$98.59_{\pm 0.04}$	$97.99_{\pm 0.10}$	$98.98_{\pm 0.07}$	$\mathbf{74.17}_{\pm 0.16}$	$98.63_{\pm 0.06}$	$94.25_{\pm 0.04}$
UniVAD + CPR	$97.04_{\pm 0.10}$	$98.57_{\pm 0.07}$	$97.88_{\pm 0.17}$	$98.91_{\pm 0.07}$	$73.99_{\pm 0.23}$	$98.67_{\pm 0.06}$	$94.18_{\pm 0.05}$
UniVAD + GRef	$\mathbf{97.16}_{\pm 0.00}$	$\mathbf{98.66}_{\pm 0.00}$	$\mathbf{98.02}_{\pm 0.00}$	$\mathbf{99.04}_{\pm 0.00}$	$74.16_{\pm 0.10}$	$\mathbf{98.68}_{\pm 0.00}$	$\mathbf{94.29}_{\pm 0.01}$

Table 7: Cross-dataset Pixel AUROC (%) with eight reference images. Results report mean ± standard deviation over three runs. Avg. is the unweighted mean across the six datasets. † denotes detectors trained with GRef-selected references. Bold indicates the best result within each group.

Method	MVTecAD	VisA	BTAD	MPDD	LOCO	MVTec3D	Avg.
PatchCore + Fixed	$84.96_{\pm 0.76}$	$82.37_{\pm 0.95}$	$71.48_{\pm 0.40}$	$87.91_{\pm 1.45}$	$61.01_{\pm 0.32}$	$88.20_{\pm 0.26}$	$79.32_{\pm 0.39}$
PatchCore + Random	$85.18_{\pm 0.17}$	$81.43_{\pm 0.45}$	$71.27_{\pm 0.37}$	$87.54_{\pm 0.33}$	$60.97_{\pm 0.24}$	$87.48_{\pm 0.14}$	$78.98_{\pm 0.20}$
PatchCore + Similarity	$88.96_{\pm 0.03}$	$82.49_{\pm 0.15}$	$71.48_{\pm 0.21}$	$87.93_{\pm 0.22}$	$61.64_{\pm 0.06}$	$88.71_{\pm 0.12}$	$80.20_{\pm 0.04}$
PatchCore + CPR	$89.95_{\pm 0.24}$	$83.42_{\pm 0.21}$	$70.64_{\pm 0.36}$	$\mathbf{91.45}_{\pm 0.20}$	$62.28_{\pm 0.45}$	$90.48_{\pm 0.13}$	$81.37_{\pm 0.06}$
PatchCore + GRef	$\mathbf{91.69}_{\pm 0.00}$	$\mathbf{85.22}_{\pm 0.14}$	$\mathbf{73.66}_{\pm 0.17}$	$91.36_{\pm 0.10}$	$\mathbf{64.06}_{\pm 0.14}$	$\mathbf{91.68}_{\pm 0.14}$	$\mathbf{82.95}_{\pm 0.10}$
RegAD + Fixed	$87.65_{\pm 0.96}$	$78.48_{\pm 1.24}$	$76.46_{\pm 1.72}$	$82.81_{\pm 2.63}$	$67.75_{\pm 1.07}$	$89.73_{\pm 0.57}$	$80.48_{\pm 0.80}$
RegAD + Random	$87.51_{\pm 1.19}$	$78.11_{\pm 2.01}$	$76.38_{\pm 0.84}$	$82.68_{\pm 2.33}$	$67.39_{\pm 0.82}$	$89.55_{\pm 0.55}$	$80.27_{\pm 0.57}$
RegAD + Similarity	$88.65_{\pm 0.70}$	$78.87_{\pm 1.50}$	$75.76_{\pm 1.01}$	$84.22_{\pm 1.78}$	$67.57_{\pm 0.77}$	$90.31_{\pm 0.56}$	$80.89_{\pm 0.51}$
RegAD + CPR	$89.54_{\pm 0.86}$	$81.32_{\pm 1.29}$	$75.32_{\pm 0.97}$	$86.35_{\pm 1.73}$	$68.99_{\pm 0.85}$	$91.62_{\pm 0.55}$	$82.19_{\pm 0.60}$
RegAD + GRef	$\mathbf{91.26}_{\pm 0.61}$	$\mathbf{83.13}_{\pm 1.70}$	$\mathbf{78.25}_{\pm 1.01}$	$87.19_{\pm 1.81}$	$\mathbf{70.36}_{\pm 0.95}$	$92.35_{\pm 0.52}$	$\mathbf{83.76}_{\pm 0.56}$
RegAD + GRef†	$90.82_{\pm 0.63}$	$82.85_{\pm 2.40}$	$78.00_{\pm 0.47}$	$\mathbf{88.57}_{\pm 0.57}$	$69.04_{\pm 1.18}$	$\mathbf{92.41}_{\pm 0.35}$	$83.61_{\pm 0.45}$
ResAD++ + Fixed	$88.86_{\pm 0.83}$	$86.36_{\pm 0.41}$	$78.55_{\pm 0.28}$	$91.88_{\pm 0.26}$	$59.29_{\pm 0.00}$	$91.16_{\pm 0.22}$	$82.68_{\pm 0.00}$
ResAD++ + Random	$88.78_{\pm 0.00}$	$85.94_{\pm 0.33}$	$78.43_{\pm 0.30}$	$91.86_{\pm 0.22}$	$59.04_{\pm 0.33}$	$90.80_{\pm 0.24}$	$82.48_{\pm 0.22}$
ResAD++ + Similarity	$90.84_{\pm 0.01}$	$86.53_{\pm 0.18}$	$78.88_{\pm 0.20}$	$92.59_{\pm 0.01}$	$59.63_{\pm 0.05}$	$91.53_{\pm 0.05}$	$83.33_{\pm 0.07}$
ResAD++ + CPR	$91.48_{\pm 0.17}$	$86.99_{\pm 0.46}$	$77.33_{\pm 0.63}$	$94.23_{\pm 0.01}$	$60.17_{\pm 0.14}$	$92.26_{\pm 0.16}$	$83.74_{\pm 0.15}$
ResAD++ + GRef	$\mathbf{92.83}_{\pm 0.00}$	$88.88_{\pm 0.10}$	$80.67_{\pm 0.00}$	$\mathbf{94.91}_{\pm 0.00}$	$61.26_{\pm 0.17}$	$\mathbf{92.91}_{\pm 0.10}$	$85.25_{\pm 0.00}$
ResAD++ + GRef†	$92.58_{\pm 0.00}$	$\mathbf{89.46}_{\pm 0.10}$	$\mathbf{81.12}_{\pm 0.00}$	$94.61_{\pm 0.00}$	$\mathbf{61.85}_{\pm 0.17}$	$92.41_{\pm 0.10}$	$\mathbf{85.34}_{\pm 0.03}$
UniVAD + Fixed	$93.41_{\pm 0.17}$	$94.92_{\pm 0.30}$	$82.69_{\pm 0.86}$	$96.35_{\pm 0.00}$	$70.45_{\pm 0.20}$	$95.86_{\pm 0.00}$	$88.95_{\pm 0.10}$
UniVAD + Random	$93.59_{\pm 0.18}$	$94.81_{\pm 0.39}$	$82.37_{\pm 0.35}$	$96.31_{\pm 0.30}$	$70.13_{\pm 0.35}$	$95.79_{\pm 0.25}$	$88.83_{\pm 0.12}$
UniVAD + Similarity	$94.06_{\pm 0.03}$	$95.21_{\pm 0.16}$	$82.79_{\pm 0.21}$	$97.41_{\pm 0.19}$	$70.96_{\pm 0.05}$	$96.01_{\pm 0.12}$	$89.41_{\pm 0.07}$
UniVAD + CPR	$93.99_{\pm 0.21}$	$94.83_{\pm 0.42}$	$82.73_{\pm 0.66}$	$97.22_{\pm 0.20}$	$70.77_{\pm 0.41}$	$96.04_{\pm 0.16}$	$89.26_{\pm 0.14}$
UniVAD + GRef	$\mathbf{94.16}_{\pm 0.00}$	$\mathbf{95.30}_{\pm 0.10}$	$\mathbf{83.07}_{\pm 0.00}$	$\mathbf{97.49}_{\pm 0.00}$	$\mathbf{71.84}_{\pm 0.10}$	$\mathbf{96.13}_{\pm 0.00}$	$\mathbf{89.66}_{\pm 0.02}$

The task loss and the empirical and population risks are

$$L_h(T) = \frac{1}{n_T} \sum_{i=1}^{n_T} \ell_h((q_i, \mathcal{R}), y_i),$$

$$\widehat{R}(h) = \frac{1}{m} \sum_{t=1}^{m} L_h(T_t), \qquad R(h) = \mathbb{E}_{T \sim \mathfrak{P}} L_h(T).$$

The parameter classes, fixed feature extractors, graph construction, reference budgets, MMR coefficient, and norm bounds are specified independently of the sampled historical tasks. The selected weights themselves may depend on those tasks. Feature and weight bounds hold uniformly over all admissible inputs and parameter values, rather than only on the observed training set. In particular, c_v bounds the detector's complete query–reference feature input, including all K references.

The quantities L_ψ and L_θ count trainable matrix transformations. A GIN message-passing layer containing a multilayer MLP contributes all of that MLP's trainable matrices. The widths d_ψ, d_θ bound both dimensions of the corresponding parameter matrices, including the input and output layers and any bias augmentation. Thus the numbers of free scalar parameters obey $P_\psi \le L_\psi d_\psi^2$ and $P_\theta \le L_\theta d_\theta^2$. Frozen parameters do not contribute to these dimensions. All trainable quantities must be included in the parameter counts and in the network sensitivity analysis.

The nondegeneracy condition concerns the raw descriptor $a_\psi(x)$, with

$$z_\psi(x) = a_\psi(x)/\|a_\psi(x)\|_2, \qquad \inf_{\psi,x} \|a_\psi(x)\|_2 \ge \nu > 0.$$

A lower bound on the already normalized descriptor would not control the sensitivity of this normalization operation.

The next lemma supplies both ingredients needed for the final concentration argument. Boundedness permits task-level concentration, and output continuity converts detector covering precision into an analysis-loss error.

Lemma 1 (Boundedness and output continuity of the loss)**.** *For $\widehat{y}, \widehat{y}', y \in [0,1]^{H\times W}$,*

$$0 \le d_{\mathrm{pix}}(\widehat{y}, y) \le 1, \qquad |d_{\mathrm{pix}}(\widehat{y}, y) - d_{\mathrm{pix}}(\widehat{y}', y)| \le d_{\mathrm{pix}}(\widehat{y}, \widehat{y}'). \tag{8}$$

Consequently, $0 \le \ell_h \le 1$ and $0 \le L_h(T) \le 1$. If two models select the same ordered reference sequence and their detector outputs differ by at most u in pixel-averaged L_1 distance, their analysis losses differ by at most u.

Proof: Every pixelwise absolute error lies in $[0,1]$. The reverse triangle inequality gives

$$\big|\|\widehat{y} - y\|_1 - \|\widehat{y}' - y\|_1\big| \le \|\widehat{y} - \widehat{y}'\|_1.$$

Division by HW proves the claims. Q.E.D.

H.2 From Network Parameters to Outputs

The finite-cover argument requires output bounds for nearby parameter values. We therefore first derive the architecture constants appearing in the theorem, before considering the discrete reference choices.

We now derive the layerwise bounds for the affine/residual networks specified in the main text. Norm bounds on weights alone would not justify them for arbitrary network operations. For each trainable layer or parameter block, let $T(H; W)$ be its state update, including its fixed intervening operations. The properties used in the proof are

$$\begin{aligned} \|T(H;W)\|_F &\le (1+ac)\|H\|_F, \\ \|T(H;W) - T(H';W')\|_F &\le (1+ac)\|H - H'\|_F \\ &\quad + a\|H'\|_F\|W - W'\|_F. \end{aligned} \tag{9}$$

These bounds are uniform over the allowed weights and states. The same a is used for all encoder and detector layers. The factor $1 + ac$ permits within-layer residual connections. The separate regional-feature residual in the encoder is treated explicitly below.

For example, take $T(H; W) = \sigma(AHW)$, where $\|A\|_2 \le a$, σ is 1-Lipschitz, and $\sigma(0) = 0$. Then

$$\|T(H;W)\|_F \le ac\|H\|_F,$$

while

$$\begin{aligned} \|T(H;W) - T(H';W')\|_F &\le \|A(H - H')W + AH'(W - W')\|_F \\ &\le ac\|H - H'\|_F + a\|H'\|_F\|W - W'\|_F. \end{aligned}$$

Adding an identity residual increases the input factor to $1 + ac$ and gives Eq. (9). An ordinary dense layer uses $A = I$. For fixed transformations before and after the trainable matrix, their combined Lipschitz and norm-growth factors must be bounded by a on the admissible domain.

Biases, when present, are represented using homogeneous coordinates. For a vector affine layer, the corresponding matrix is

$$\widetilde{W} = \begin{pmatrix} W & b_0 \\ 0 & 1 \end{pmatrix}, \qquad \widetilde{v} = \begin{pmatrix} v \\ 1 \end{pmatrix}.$$

Here b_0 is a bias vector, distinct from the common norm bound b. Weight-norm and dimension bounds apply to these augmented matrices; input bounds include the constant coordinates. Fixed entries of the augmented matrices do not vary when parameters are compared. Their inclusion only enlarges the ambient covering bound. In a homogeneous representation, the activation extends as $(h, t) \mapsto (\sigma(h), t)$ and is still 1-Lipschitz and zero-preserving. A residual addition acts on the feature coordinates while retaining the homogeneous coordinate once, which does not increase the bounds above. Fixed transformations and their nonexpansiveness are understood on the full augmented states. A bias-free network is a special case requiring no augmentation.

We first propagate the layerwise bounds through a chain of trainable transformations. Products of layer spectral norms are also used to control network amplification in norm-based generalization analyses (Bartlett et al., 2017). The following recurrence gives the constants needed for our graph aggregation and residual architecture.

Lemma 2 (Continuity of a chain of trainable layers)**.** *Consider a chain of $L \geq 1$ layers satisfying Eq.* (9)*, with a common input v obeying $\|v\|_F \leq r$. Set $t = 1 + ac$ for this proof. Its output norm is at most rt^L, its input Lipschitz factor is at most t^L, and its parameter variation satisfies*

$$\|F_W(v) - F_{W'}(v)\|_F \leq art^{L-1} \sum_{l=1}^{L} \|W_l - W_l'\|_F. \tag{10}$$

Consequently, its parameter Lipschitz factor is at most $ar\sqrt{L}\, t^{L-1}$ for concatenated parameter vectors.

Proof: The norm and input Lipschitz bounds follow by induction using Eq. (9). For the parameter comparison, let $e_l = \|H_l - H_l'\|_F$, with $e_0 = 0$. Since $\|H_{l-1}'\|_F \leq rt^{l-1}$,

$$e_l \leq te_{l-1} + art^{l-1}\|W_l - W_l'\|_F.$$

Repeated substitution gives

$$e_L \leq \sum_{l=1}^{L} t^{L-l} art^{l-1}\|W_l - W_l'\|_F = art^{L-1} \sum_{l=1}^{L} \|W_l - W_l'\|_F.$$

Finally,

$$\sum_{l=1}^{L} \|W_l - W_l'\|_F \leq \sqrt{L} \left(\sum_{l=1}^{L} \|W_l - W_l'\|_F^2 \right)^{1/2}.$$

Q.E.D.

We now apply Lemma 2 to the encoder's GIN branch and projection MLP, retaining the regional-feature residual connection. The same lemma controls the detector for a fixed ordered reference sequence. These estimates will determine the radii of the finite parameter covers used later.

Lemma 3 (Parameter continuity of the encoder and detector)**.** *With the explicit constants in Theorem 1,*

$$\sup_x \|z_\psi(x) - z_{\psi'}(x)\|_2 \leq \kappa_\psi \|\psi - \psi'\|_2, \tag{11}$$

$$\sup_{q,\mathcal{S}} \frac{\|f_\theta(q, \mathcal{S}) - f_{\theta'}(q, \mathcal{S})\|_1}{HW} \leq \kappa_\theta \|\theta - \theta'\|_2. \tag{12}$$

Here $\mathcal{S}$ ranges over ordered sequences of K references, and parameter vectors concatenate the variable entries of their parameter matrices.

Proof: Write $L_\psi = L_G + L_P$, where $L_G, L_P \geq 1$ are the numbers of trainable layers in the GIN branch and projection MLP. Let G_{ψ_G} and P_{ψ_P} denote these two components, and set $t = 1 + ac$. For the same input $v = v_\psi(x)$, write

$$w = \mathrm{vec}[v + G_{\psi_G}(v)], \qquad w' = \mathrm{vec}[v + G_{\psi'_G}(v)].$$

Lemma 2 yields

$$\|w'\|_2 \leq c_v(1 + t^{L_G}),$$
$$\|w - w'\|_2 \leq ac_v t^{L_G - 1} \sum_{l \in G} \|W_{\psi,l} - W'_{\psi,l}\|_F.$$

If homogeneous coordinates are used, the fixed residual addition preserves its constant coordinates and satisfies the same norm and difference bounds. Split the raw encoder output difference as

$$\begin{aligned}\|a_\psi(x) - a_{\psi'}(x)\|_2 &\leq \|P_{\psi_P}(w) - P_{\psi_P}(w')\|_2 \\ &\quad + \|P_{\psi_P}(w') - P_{\psi'_P}(w')\|_2 \\ &\leq ac_v t^{L_\psi - 1} \sum_{l \in G} \|W_{\psi,l} - W'_{\psi,l}\|_F \\ &\quad + ac_v(1 + t^{L_G}) t^{L_P - 1} \sum_{l \in P} \|W_{\psi,l} - W'_{\psi,l}\|_F.\end{aligned}$$

Since $t \geq 1$, $1 + t^{L_G} \leq 2t^{L_G}$. Therefore,

$$\begin{aligned}\|a_\psi(x) - a_{\psi'}(x)\|_2 &\leq 2ac_v t^{L_\psi - 1} \sum_{l=1}^{L_\psi} \|W_{\psi,l} - W'_{\psi,l}\|_F \\ &\leq 2ac_v \sqrt{L_\psi}\, t^{L_\psi - 1} \|\psi - \psi'\|_2.\end{aligned}$$

This accounts for the encoder's regional-feature residual connection. For vectors v, w whose norms are at least ν,

$$\left\| \frac{v}{\|v\|_2} - \frac{w}{\|w\|_2} \right\|_2^2 = \frac{\|v - w\|_2^2 - (\|v\|_2 - \|w\|_2)^2}{\|v\|_2 \|w\|_2} \leq \frac{\|v - w\|_2^2}{\nu^2}.$$

Applying this to the raw descriptors gives Eq. (11).

For the detector, its complete fixed feature input has norm at most c_v uniformly over the admissible query–reference inputs. Lemma 2 gives a parameter Lipschitz factor $ac_v\sqrt{L_\theta}\, t^{L_\theta - 1}$ in output Euclidean norm. Its final fixed 1-Lipschitz output map does not increase this factor. Finally, for any output difference $v \in \mathbb{R}^{HW}$,

$$\frac{\|v\|_1}{HW} \leq \frac{\|v\|_2}{\sqrt{HW}}.$$

This proves Eq. (12). All bounds hold uniformly over admissible inputs. Q.E.D.

H.3 From Descriptor Perturbations to Stable Selection

Lemma 3 controls descriptor changes, but descriptor continuity alone does not ensure identical references. We next identify the score gaps that preserve Top-M retrieval and each MMR decision. This will allow a covered model to use the same ordered references whenever the selection margin is sufficiently large.

For a query–pool pair $e = (q, \mathcal{R})$, write

$$s_\psi(x, x') = \frac{1 + z_\psi(x)^\top z_\psi(x')}{2}, \qquad g_{i,\psi} = s_\psi(q, r_i).$$

Let $\mathcal{C}_\psi(e)$ be the Top-M candidate set, and let $g_{(1),\psi} \geq \cdots \geq g_{(N),\psi}$ be the ordered scores. The retrieval margin is

$$\Delta_{\mathrm{ret},\psi}(e) = g_{(M),\psi} - g_{(M+1),\psi}.$$

At MMR step j, let $S_{j-1,\psi}$ be the previously selected set and define, for each remaining candidate,

$$h_{j,\psi}(i) = \lambda g_{i,\psi} - (1-\lambda) \max_{r \in S_{j-1,\psi}} s_\psi(r_i, r), \qquad 0 \le \lambda \le 1.$$

The maximum over the empty set is defined to be zero here. Let $i_{j,\psi}$ be the selected index, with a fixed deterministic rule for breaking ties. Define

$$\Delta_{j,\psi}(e) = h_{j,\psi}(i_{j,\psi}) - \max_{i \in \mathcal{C}_\psi(e) \setminus (S_{j-1,\psi} \cup \{i_{j,\psi}\})} h_{j,\psi}(i).$$

In this expression, sets of references and their index sets are identified. Since $K < M$, the runner-up set is nonempty at every step. Finally, define

$$\Gamma_\psi(e) = \min \left\{ \Delta_{\mathrm{ret},\psi}(e), \Delta_{1,\psi}(e), \ldots, \Delta_{K,\psi}(e) \right\}. \tag{13}$$

A tie at any relevant boundary gives a zero margin.

For a fixed threshold $\gamma > 0$, define the empirical small-margin frequency by

$$\widehat{p}_\psi(\gamma) = \frac{1}{m} \sum_{t=1}^{m} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{1}\{\Gamma_\psi(q_{t,i}, \mathcal{R}_t) \le \gamma\}. \tag{14}$$

This is the fraction of small-margin queries within each task, averaged equally over the m historical tasks. Its weighting matches that of $\widehat{R}(\psi, \theta)$. When all n_t are equal, it also equals the fraction among all historical queries. A small margin does not assert that the selected sequence changes; it marks a query for which the margin-based stability guarantee below may not apply. The term $\widehat{p}_\psi(\gamma)$ will enter the risk bound when we average the pointwise small-margin indicator in the loss comparison in Eq. (20).

Lemma 4 (Stability of the selection sequence)**.** *Suppose* $\sup_x \|z_\psi(x) - z_{\psi'}(x)\|_2 \le \eta$. *If* $\Gamma_\psi(e) > 2\eta$*, both encoders return the same Top-*M *candidate set and ordered sequence of* K *references. Moreover, whenever these candidate sets and selection sequences agree,*

$$|\Gamma_\psi(e) - \Gamma_{\psi'}(e)| \le 2\eta. \tag{15}$$

Proof: Since all descriptors have unit norm,

$$\begin{aligned} |s_\psi(x, x') - s_{\psi'}(x, x')| &\le \tfrac{1}{2} \|z_\psi(x) - z_{\psi'}(x)\|_2 \\ &\quad + \tfrac{1}{2} \|z_\psi(x') - z_{\psi'}(x')\|_2 \le \eta. \end{aligned}$$

Thus each retrieval score changes by at most η. A boundary gap greater than 2η preserves the candidate set.

Suppose inductively that the previous MMR selections agree. The redundancy scores then compare maxima over the same reference set, so each changes by at most η. Every MMR objective changes by at most $\lambda\eta + (1-\lambda)\eta = \eta$. A winner–runner-up gap greater than 2η therefore preserves the next selected reference. Induction proves sequence stability.

Each order statistic of the retrieval scores changes by at most η, so the retrieval gap changes by at most 2η. When the selection sequences agree, each MMR winner is the same and its runner-up maximum is taken over the same remaining indices. Its gap consequently changes by at most 2η. Taking the minimum of these gaps proves Eq. (15). Q.E.D.

H.4 Finite Covers of the Network Classes

We can now replace the continuous parameter classes by finitely many representatives. Lemma 3 converts parameter covering radii into uniform output errors. We choose encoder precision $\gamma/4$ so that Lemma 4 can also transfer selection margins between a model and its representative.

Lemma 5 (Architecture-dependent covering numbers)**.** *For any* $\gamma, u > 0$*, there are finite parameter sets* $\overline{\Psi} \subset \Psi$ *and* $\overline{\Theta} \subset \Theta$ *such that every* (ψ, θ) *has a representative* $(\bar{\psi}, \bar{\theta})$ *satisfying*

$$\sup_x \|z_\psi(x) - z_{\bar{\psi}}(x)\|_2 \le \gamma/4, \tag{16}$$

$$\sup_{q,\mathcal{S}} \frac{\|f_\theta(q, \mathcal{S}) - f_{\bar{\theta}}(q, \mathcal{S})\|_1}{HW} \le u. \tag{17}$$

Their cardinalities satisfy

$$\log\left(|\overline{\Psi}|\,|\overline{\Theta}|\right) \le \mathcal{C}(\gamma, u). \tag{18}$$

Proof: The encoder parameter vector has dimension $P_\psi \le L_\psi d_\psi^2$ and satisfies

$$\|\psi\|_2^2 \le \sum_{l=1}^{L_\psi} \|W_{\psi,l}\|_F^2 \le L_\psi b^2.$$

The detector similarly has dimension $P_\theta \le L_\theta d_\theta^2$ and radius $b\sqrt{L_\theta}$.

We use the standard volumetric covering argument (Vershynin, 2026, Section 4.3). Any subset of a radius-A ball in $\mathbb{R}^P$ admits a Euclidean r-net with centers in that subset and cardinality at most $(1+2A/r)^P$. Indeed, balls of radius $r/2$ around a maximal r-separated subset have disjoint interiors and lie in a ball of radius $A + r/2$; comparing volumes gives the bound, and maximality ensures the covering property.

Apply this fact with

$$r_\psi = \frac{\gamma}{4\kappa_\psi}, \qquad r_\theta = \frac{u}{\kappa_\theta}.$$

Lemma 3 gives Eqs. (16)–(17), and the volume bounds give

$$\log|\overline{\Psi}| \le L_\psi d_\psi^2 \log\left(1 + \frac{8b\sqrt{L_\psi}\kappa_\psi}{\gamma}\right),$$

$$\log|\overline{\Theta}| \le L_\theta d_\theta^2 \log\left(1 + \frac{2b\sqrt{L_\theta}\kappa_\theta}{u}\right).$$

The covers are fixed independently of the historical tasks. Their centers remain in the admissible classes, including the encoder nondegeneracy constraint. Q.E.D.

H.5 Completion of the Generalization Proof

The preceding lemmas provide finite model representatives and a condition for preserving their reference choices. It remains to handle inputs with small margins and obtain a simultaneous risk bound. We construct a bounded loss envelope for each representative, apply concentration across independent tasks, and then transfer the result to every admissible model.

Proof: [Proof of Theorem 1] Write $h = (\psi, \theta)$ and let $\bar{h} = (\bar{\psi}, \bar{\theta})$ be a representative supplied by Lemma 5. For $e = (q, \mathcal{R})$, define the representative's upper envelope

$$U_{\bar{h}}(e, y) = \min\left\{1, \ell_{\bar{h}}(e, y) + u + \mathbf{1}\{\Gamma_{\bar{\psi}}(e) \le \gamma/2\}\right\}. \tag{19}$$

All losses and envelopes are measurable and lie in $[0, 1]$. On sufficiently large-margin inputs, the representative selects the same references, and its detector error can be controlled by u. On small-margin inputs, we instead use the unit upper bound on the loss. The following comparison makes this distinction explicit and identifies the source of the small-margin penalty. We first prove the pointwise bounds

$$\ell_h(e, y) \le U_{\bar{h}}(e, y) \le \ell_h(e, y) + 2u + \mathbf{1}\{\Gamma_\psi(e) \le \gamma\}. \tag{20}$$

For the first inequality, if $\Gamma_{\bar{\psi}}(e) \le \gamma/2$, then $U_{\bar{h}} = 1$. Otherwise, the encoder discrepancy is at most $\gamma/4$, so Lemma 4, applied with $\bar{\psi}$ as the initial encoder, implies identical ordered reference sequences. Lemma 1 and Eq. (17) give

$$|\ell_h(e, y) - \ell_{\bar{h}}(e, y)| \le u.$$

Since also $\ell_h \le 1$, this proves $\ell_h \le U_{\bar{h}}$.

For the second inequality, if $\Gamma_\psi(e) \le \gamma$, it follows directly from $U_{\bar{h}} \le 1$. Otherwise, Lemma 4 gives identical sequences and

$$\Gamma_{\bar{\psi}}(e) \ge \Gamma_\psi(e) - \gamma/2 > \gamma/2.$$

The indicator in Eq. (19) vanishes, and hence

$$U_{\bar{h}}(e, y) \le \ell_{\bar{h}}(e, y) + u \le \ell_h(e, y) + 2u.$$

This establishes Eq. (20).

For each representative define the task-level envelope

$$G_{\bar{h}}(T) = \frac{1}{n_T} \sum_{i=1}^{n_T} U_{\bar{h}}((q_i, \mathcal{R}), y_i).$$

Each $G_{\bar{h}}(T)$ lies in $[0, 1]$.

Averaging Eq. (20) over the queries within a task gives

$$\begin{aligned} L_h(T) \le G_{\bar{h}}(T) \le L_h(T) + 2u \\ + \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbf{1}\{\Gamma_\psi(q_i, \mathcal{R}) \le \gamma\}. \end{aligned} \tag{21}$$

Averaging this upper bound over the historical tasks yields

$$\begin{aligned} \frac{1}{m} \sum_{t=1}^{m} G_{\bar{h}}(T_t) &\le \widehat{R}(\psi, \theta) + 2u \\ &\quad + \frac{1}{m} \sum_{t=1}^{m} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{1}\{\Gamma_\psi(q_{t,i}, \mathcal{R}_t) \le \gamma\} \\ &= \widehat{R}(\psi, \theta) + 2u + \widehat{p}_\psi(\gamma). \end{aligned} \tag{22}$$

The last equality follows from Eq. (14). Thus, $\widehat{p}_\psi(\gamma)$ arises from averaging the pointwise penalties for small-margin queries, using the same task and query weights as $\widehat{R}(\psi, \theta)$. We next combine this empirical comparison with a concentration bound to obtain the final risk bound.

It remains to relate the empirical envelope to its expectation. Let $J = |\overline{\Psi}|\,|\overline{\Theta}|$. For each fixed representative and $\varepsilon > 0$, Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr\left[\mathbb{E}G_{\bar{h}}(T) - \frac{1}{m} \sum_{t=1}^{m} G_{\bar{h}}(T_t) > \varepsilon\right] \le e^{-2m\varepsilon^2}.$$

A union bound over the J fixed representatives, with $\varepsilon = \sqrt{\log(J/\delta)/(2m)}$, implies that, with probability at least $1 - \delta$, simultaneously for every representative,

$$\mathbb{E}_{T \sim \mathfrak{P}} G_{\bar{h}}(T) \le \frac{1}{m} \sum_{t=1}^{m} G_{\bar{h}}(T_t) + \sqrt{\frac{\log J + \log(1/\delta)}{2m}}.$$

Independence is required only across the m tasks, since each task contributes one bounded random variable. The event is established for the fixed finite covers before a data-dependent model is selected.

For any admissible $h = (\psi, \theta)$ and its representative $\bar{h}$, the pointwise domination $\ell_h \le U_{\bar{h}}$ implies $R(\psi, \theta) \le \mathbb{E}_{T \sim \mathfrak{P}} G_{\bar{h}}(T)$. On the simultaneous concentration event, we therefore obtain

$$\begin{aligned} R(\psi, \theta) &\le \mathbb{E}_{T \sim \mathfrak{P}} G_{\bar{h}}(T) \\ &\le \frac{1}{m} \sum_{t=1}^{m} G_{\bar{h}}(T_t) + \sqrt{\frac{\log J + \log(1/\delta)}{2m}} \\ &\le \widehat{R}(\psi, \theta) + \widehat{p}_\psi(\gamma) + 2u + \sqrt{\frac{\log J + \log(1/\delta)}{2m}} \\ &\le \widehat{R}(\psi, \theta) + \widehat{p}_\psi(\gamma) + 2u + \sqrt{\frac{\mathcal{C}(\gamma, u) + \log(1/\delta)}{2m}}. \end{aligned} \tag{23}$$

The third inequality uses Eq. (22), which introduces $\widehat{p}_\psi(\gamma)$ through the empirical small-margin frequency. The last inequality follows from $\log J \le \mathcal{C}(\gamma, u)$ in Eq. (18). This proves Eq. (6) simultaneously for all admissible (ψ, θ), including parameters selected using the historical tasks. Q.E.D.

H.6 Additional Remarks

Training and analysis losses. The theorem is uniform over the admissible encoder and detector parameters. It therefore applies to parameters produced by any measurable training procedure, including training with focal loss or additional auxiliary objectives, as long as the returned parameters belong to the specified classes. It does not assume minimization of the empirical L_1 risk, and does not assert that a small focal training loss implies a small empirical L_1 risk. The numerical constants in the theorem concern the L_1 analysis loss. Applying the argument to another loss requires its own boundedness and output Lipschitz constants.

Other network operations. The bounds above require each trainable block, including its fixed intermediate operations, to satisfy Eq. (9). If an implementation uses expansive normalization or other fixed operations, the relevant combined factors must instead be incorporated in a larger a before applying the displayed architecture constants. The final detector output map needs only the stated 1-Lipschitz property and bounded range; it need not vanish at zero because it is not followed by another trainable layer. A hard decision based on trainable detector features does not in general satisfy this condition; bounded matrix norms alone cannot justify it. The Top-M/MMR selector is handled separately by its decision margins.

Choice of analysis parameters and scope. The choices γ and u are fixed before applying the probability bound. To choose them from a finite grid using the same historical tasks, an additional union bound over that grid is required. The guarantee is for expected risk over an independent new task from $\mathfrak{P}$. It does not assert the same bound for every individual target product or for a different task distribution.